# DOCUMENT RESUME

| | |
|---|---|
| ED 166 245 | TM 008 266 |

| | |
|---|---|
| AUTHOR | Bossone, Richard M., Ed. |
| TITLE | The Second National Conference on Testing: Major Issues. Proceedings. |
| INSTITUTION | City Univ. of New York, N.Y. Center for Advanced Study in Education. |
| SPONS AGENCY | CTB / McGraw-Hill, Monterey, Calif. |
| REPORT NO | CASE-27-78 |
| PUB DATE | 78 |
| NOTE | 151p.; Proceedings of the National Conference (2nd, San Francisco, California, September 21-22, 1978) |
| EDRS PRICE | MF-$0.83 HC-$8.69 Plus Postage. |
| DESCRIPTORS | Conference Reports; *Educational Testing; Evaluation Methods; Federal Regulation; *Government Role; Information Dissemination; *Minimum Competency Testing; Program Evaluation; State Programs; Student Testing; *Testing Problems; *Testing Programs; *Test Interpretation; Test Results |

ABSTRACT

The speeches presented at the second national conference on testing deal with models for program evaluation, reporting of test results, minimum competency programs, and the role of state and federal government in educational testing. Various approaches to program evaluation, and its relationship to testing are described by Michael Scriven, Lee J. Cronbach, and Robert E. Stake. Responsible reporting about testing is then discussed by Frank J. Macchiarola, chancellor, New York City Public Schools; Harry Handler, deputy superintendent, Los Angeles School District; and Ronald D. Moskowitz, education correspondent, San Francisco Chronicle. The problems and prospects of minimum competency testing programs are reviewed by Chris Pipho, Walt Haney, and Ken Loewe, who describes recent experiences in Florida. Speakers on the government role in testing include Alexander I. Law, representing California; A. Craig Phillips, from North Carolina; and John W. Evans and David M. Shoemaker of the U.S. Office of Education. Concluding remarks are made by Michael Scriven. (GDC)

***********************************************************************
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
***********************************************************************

ED166245

U S DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

TM008 266

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Richard M. Bossone

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

# Proceedings

# The Second National Conference on Testing: Major Issues

September 21 & 22, 1978
San Francisco, California

Edited by
Richard M. Bossone

Center for Advanced Study in Education
The Graduate School and University Center
of the City University of New York

CASE 27-78
1978

## TABLE OF CONTENTS

Page

OPENING REMARKS
Alexander J. Burke . . . . 1

PART I EVALUATING EDUCATIONAL PROGRAMS: THE BEST MODELS AND THEIR RELATION TO TESTING

Two Main Approaches to Evaluation
Michael Scriven . . . . . 5

Evaluation in a Context of Accommodation
Lee J. Cronbach . . . . . 21

The Validity of District-Wide Test Means
Robert E. Stake . . . . . 32

PART II RESPONSIBLE REPORTING ABOUT TESTING: SUPERINTENDENTS AND THE PRESS

Testing and Its Audiences: Contexts for Reform
Frank J. Macchiarola . . . 44

Responsible Reporting About Testing: A Deputy Superintendent's Perspective
Harry Handler . . . . . . 50

Responsible Reporting About Testing: An Education Correspondent's Perspective
Ronald D. Moskowitz . . . 56

PART III MINIMUM COMPETENCY TESTING: PROBLEMS AHEAD

Minimum Competency Testing: A National Perspective of Problems Ahead
Chris Pipho . . . . . . . 65

Competency Testing: The Florida Experience
Ken Loewe . . . . . . . . 79

Problems and Prospects in Minimum Competency Testing
Walt Haney . . . . . . . . 92

Page

PART IV STATE AND FEDERAL ROLES IN TESTING: RISE OR DECLINE

The State Role in Testing: As Viewed by a California Evaluator
Alexander I. Law . . . . 106

State and Federal Roles in Testing: As Viewed by the State Superintendent of North Carolina
A. Craig Phillips . . . . 112

State and Federal Roles in Testing: As Viewed by USOE Officers
John W. Evans and David M. Shoemaker . . . . 123

SUMMATION

Themes and Counterpoint
Michael Scriven . . . . . 134

# OPENING REMARKS

Alexander J. Burke
President
McGraw-Hill Book Company

As president of the McGraw-Hill Book Company, I am pleased to open the Second National Conference on Testing, jointly sponsored by CTB/McGraw-Hill and the Center for Advanced Study in Education of the City University of New York.

The McGraw-Hill Book Company, the largest educational publisher in the world, long ago determined that involvement in education required involvement in testing. As a result of this involvement, CTB/McGraw-Hill is today the largest school-test publisher in the country and has a continuous means of learning what test publishers have always known, that tests and instructional materials are essential parts of a single learning process and must, therefore, at least in a broad sense, be closely integrated.

I mention this relationship between textbook and test publishing because both are deeply related to the substantial issues of this conference: evaluating educational programs, responsible reporting about testing, minimum competencies, and federal and state roles in education. The impact of this conference is one that affects both testing and education, both test publishing and book publishing.

Because testing is an issue that quickly divides people, because it acts as a litmus test that reveals one's basic assumptions about educational standards, accountability, trust in measurement, all of us must be concerned with the public relations impact of testing. I am concerned, as I am sure all of you are, that we find ways to assure that the public perception of the values of testing is that tests are a positive tool to reveal successes, and not just to highlight failures, a tool to guide teachers to achieve greater success with students. Testing is not the bearer of bad tidings; it is the messenger of accurate news with a great potential for setting directions for growth and improvement.

I was on a panel recently with Jesse Jackson, Director of Operation Push. He was commenting as a black leader on his views of minimum competency testing. He said that blacks had long experienced discrimination in testing, first in the use of blood tests and later, in literacy tests; and, now, one might expect some skepticism about minimum competency testing. But, he went on to say, that his view of minimum competency testing depended on an ability to trust the people giving the tests and the aims they had in mind. He was looking for the assurance that positive routes to success would be the outcome of such testing and not just some further negative reinforcement.

I think that Mr. Jackson's comments get at a fundamental issue: the contrast between the sophisticated expertise required in test development and the amateurish use and reporting of test scores. We must find new ways of developing informed public understanding of tests and their positive values. We are living in a society which daily increases the ways in which objective measures are to be used as educational or employment credentials. This will hardly be successful until and unless we establish an informed public of parents, employers, and citizens who understand testing, see its positive values as an essential part of quality education, and are willing to support the increased funding for tests that goes well beyond the present level of $1.50 per child out of the more than $2,000 we spend on educating each child in this country.

With the Center for Advanced Study in Education of the City University of New York, I welcome you to the stimulation of a very exciting group of professionals, exchanging views in a format that we hope will sharpen issues and focus disagreement. Our conference chairman, Dr. Richard Bossone, has arranged a program and a format that is worthy of that marvelous comment made by Socrates when, in the last hour of his life, he and his friends discoursed on the immortality of the soul: "The partisan is anxious only to convince his hearers. I am seeking to convince myself: to convince my hearers is a secondary matter with me .... Agree with me if I seem to you to speak the truth; or, if not, withstand me, might and main, that I may not deceive you as well as myself in my desire and like the bee leave my sting in you before I die. And now let us proceed."

# PART I - EVALUATING EDUCATIONAL PROGRAMS: THE BEST MODELS AND THEIR RELATION TO TESTING

# TWO MAIN APPROACHES TO EVALUATION

Michael Scriven
Director
Evaluation Institute
University of San Francisco

My aim here is to outline two main approaches to evaluation in general and the evaluation of educational programs in particular. The difference between these approaches is tremendously important for the future of testing, that is, its scientific future and its political future. People in testing traditionally have not taken sides in discussions about evaluation. I think that will have to change. Neutrality on the part of test experts led to overemphasis on norm referenced testing, failure to develop functional literacy tests, overemphasis on grade-equivalent scores, and other less-than-ideal situations. So perhaps it is fitting to open a conference such as this with a discussion of the general framework into which testing must fit, the framework of program evaluation.

The most common model for evaluating educational programs is what one might call the Traditional Social Science (TSS) model. In fact, many leading evaluators find it difficult even to conceive of the existence of another approach. Recent review articles have appeared with titles, such as "The State of the Art in Educational Evaluation" in which distinguished authors have demonstrated by two blunders the characteristic acute occupational myopia of TSS fans. First, they assume without a word of justification that educational evaluation can be equated with educational program evaluation, thereby dismissing all evaluation of teaching, educational administration, educational research, educational policy, educational proposals and products, curricula, tests, students, and indeed all educational evaluations. That egregious blunder is, however, not our primary concern here today since we usually look at testing in a context of program evaluation. Second, they assume (or assert) that evaluation of educational programs is to be conducted by use of TSS methodology. And that mistake is relevant to our topic. If we can show it is a mistake, it follows that we should not focus on the TSS model alone in our discussions at this conference.

The best models of program evaluation in general, and educational program evaluation in particular, are models that have been developed within education. Because there is no time here to go into a survey of these models, I shall confine myself to delineating the main dimensions of educational program evaluation. This attempt, incidentally, can rather easily be applied to educational product evaluation and one of two other types of educational evaluations. It cannot be so usefully applied, for example, to the evaluation of teachers, teaching, or student work. In giving this outline, I will say a word or two about the extent to which TSS methodology does or does not provide an adequate basis for coping with the analysis of each dimension.

This may be the best point to distinguish between the TSS approach or perspective and TSS models of evaluation. The TSS approach is a methodological one; it involves the claim that proper methods for evaluation are those of TSS. But a model must give us some idea as to how we should apply those methods and assemble the results. In this respect, those favoring the TSS approach have traditionally had an extremely simplistic idea: that program evaluation amounts to determining whether the program goals have been achieved. It is obvious from this idea where testing enters. There have been two improvements on that grossly inadequate model, which we can call the goal-achievement model. These improvements have been more honored in the description than the practice of TSS evaluation, and that they occur at all is partially due to their strong endorsement by opponents of the TSS approach. They do not bear on the fact that the TSS model is a goal-based model of program evaluation and that is, in and of itself, a fatal flaw. For, although there are certainly some circumstances in which evaluations should be related to the goals of programs, there are certainly other circumstances in which the evaluation, or a large part of it, should be done without knowledge of or reference to those goals. Goal-based evaluation is a management serving type of evaluation; a client or consumer-oriented evaluation, for example, does not need, and is severely contaminated by, a knowledge of goals. Goal-free evaluation is as essential in program evaluation as double-blind methodology is in drug evaluation. Thus, the TSS models of evaluation are massively biased toward the management point of view.

The particular relevance of our discussion of the TSS approach to testing is two-fold. In the first place, most testing specialists were originally trained in TSS methods and hence they very naturally and commonly accept the TSS approach to program evaluation. It seems, after all, a respectable and scientific approach. And testing and instruction occupies a much larger place, relatively speaking, in the TSS approach than in the alternative I am supporting. Second, testing does, of course, have a place of honor in any listing of the dimensions of program evaluation, and it should be of specific interest to see what has to be included besides testing.

The conception of evaluation which lies behind the checklist I am about to present could be described as the view that evaluation is an extremely complex and radically different new discipline; this is most unlike TSS in its total complexion yet one which requires many of the TSS methods. On this basis, evaluation includes much of TSS but goes beyond it thereby implying that evaluation may in some sense be a more general model than that of the traditional social sciences. And I will endorse that claim; I think that, far from TSS being all there is to evaluation or even the best model for it, evaluation--properly understood--is the best model for the social sciences, a much better model than TSS ever provided. We need not go into that debate here, but perhaps it should be borne in mind that more is at stake than the relatively limited question of identifying the dimensions of evaluation; there is also the possibility that we are outlining a new conception of social sciences, of which TSS is merely a special case.

Let me suggest, then, that the following dimensions or components are typically involved--to varying degrees--in educational program evaluation. We can refer to this approach as the Complex New Discipline (CND) conception of evaluation. Incidentally, it should be stressed that the CND approach is completely neutral as between a goal-based or goal-free orientation. It can be used equally well for either, though I think it should involve a goal-free approach.

There is another crucial feature of the CND model. Because it is grounded in a conception of evaluation which views program evaluation as one very limited case, it inevitably leads to the realization that evaluations also have to be evaluated, an explicitly self-critical stance expressed in the maxim that evaluation begins at home; that is, the evaluator must practice what she/he preaches. Hence, virtually every consideration listed here as relevant to the evaluation of a program must, for reasons of professional propriety as well as scientific necessity, be applied to the evaluation of that program. We call that the self-referential dimension of an evaluation or the metaevaluation and list it in the appropriate place below: it generates a second column of analysis for every evaluation.

Finally, we must stress that the following list is not diachronic but iterative. We may work through it in the order given or in that order as far as is feasible, but we will always have to go through it again and again, modifying our treatment of the earlier points in light of later discoveries. Many are mutually dependent, and cannot be completed before we pass on; they can be only approximated. Now to the list.

_Metaevaluation_. This is an important dimension of each evaluation, which focuses on such questions as its cost effectiveness, the propriety of its processes, etc. TSS tends to ignore this because of the basic tendency in TSS to see the experimenter as a separate entity from the subject and, in particular, to see evaluation as program evaluation not as evaluation of a wide range of entities including evaluation.

_Context_. This involves the history of the program, rationale, goals for it (provided the evaluation is not goal-free), staff, clients, political climate and context, audiences, etc. TSS evaluation will often begin with this, but will rarely recognize the fact that CND attaches great importance to it, because in many cases that may be most of an evaluation. Historians do evaluations, and they use none of the apparatus of TSS. Hence, there will be times when programs are evaluated historically, qualitatively, and still objectively. One might say that the TSS approach sees the history of the project as part of the introduction to the evaluation; whereas, the CND approach considers the possibility that it is the whole thing.

Although we will not repeat this exercise with each of the checkpoints, we will indicate the way in which the metaevaluation "parallel process" operates; in this case, there would be a discussion of the history of the evaluation of the program, a discussion of its rationale, and of the political climate in which it has been commissioned. The importance of this in doing evaluation is that it focuses the attention of the evaluator on possibly serious biases or expectations; in addition, it is good scientific practice.

Subject. What is it that is being evaluated? Again, a seemingly innocuous checkpoint which has grown increasingly important as we learn from experience. Almost all apparent descriptions of programs are in fact quasi-evaluative, for example, "functional literacy programs" (or test). It may be extremely important to try to develop alternative descriptions, to look at the origin of the descriptions offered, to note discrepancies between them, and to note neutral language in the descriptions. In fact, there are extreme cases where almost the entire evaluation involves getting the description right. The TSS method is not inclined to recognize the power of description as such; on the contrary, it treats description as just a prolegomenon. Other and important aspects of the description, which both CND and TSS are likely to cover, include such matters as entry skill requirements and support system arrangements.

Need. The needs assessment is the key place where values get into evaluation. This is an entirely new dimension of evaluation in contrast with TSS investigations. TSS has produced models for needs assessment which are inappropriate, based on completely incorrect conceptual analysis of need and a desperate attempt to equate it to the results of a survey or testing program. There are even textbooks on the market in which children's need for remedial reading is said to be demonstrated by showing they are reading below grade-equivalent levels. The crudity of this conception of need is unequalled in the list of oversimplifications for which TSS is responsible. Such an approach has had very serious social consequences. And, of course, it is not at all essential that a careful social scientist should make such a mistake; it just happens to be the case that TSS evaluations have tended to make it. The reason for this has to be sought in the extreme anxiety about the value element in social science which is part of the traditional concept of TSS. The CND position is if you have to get your needs assessment from a value-laden context riddled with ethical discussions and analysis of traditionally controversial areas, such as "the obligations of the school system," then that is where you must get it. One might as well roll up one's sleeves, learn how to include ethical or analytical discussions or latent function theory, and get the job done.

There is no assumption in the CND approach that one will be able to come up with simple answers; but there is the belief that one has to come up with the best answers or the best proof that several answers are equally good in order to get the job done. In short, evaluators cannot ignore the value judgment factor.

Market. Traditionally this has been identified with need; but CND evaluators, who possess much greater sensitivity to differences between wants and needs, are more likely to avoid the blunder of identifying a survey of wants with a methodology of needs assessment. The importance of market research is still great; one does not do any good by creating a program or product for which there is no market. An evaluation must be realistic in the sense of evaluating the worth of something (product or program) in terms of what good it is actually going to do, not just in terms of what good it might do.

Effects. Here we come to the first of a number of checkpoints where TSS methods are excellent, and can be applied "right out of the box" (or with the assistance of some enlightened developments like quasi-experimental design). The pitfall here for TSS evaluators is that they have normally suffered from severe tunnel vision, looking for effects only in the direction of the goals of the program. Of course, side effects may be more important than intended effects and must be sought systematically and assiduously. The only specific methodology for identifying side effects, of which I am aware, is the use of the goal-free mode, which most TSS evaluators do not employ. And it should be noted that this checkpoint is, of course, where testing activities come into full play.

Causation. What are the effects the effects of? Here again the TSS approach is well conceived although somewhat limited. The necessity for introducing the tools of the detective, historian, and pathologist is sometimes evident from the impossibility of establishing either control groups or quasi-control groups. And those disciplines are not part of the TSS package. Hence, the development of the *modus operandi* method, which is not, as one might expect, a TSS contribution. In tricky cases, TSS evaluators do run into difficulty with causation because of weak conceptual analysis; they usually think that the absence of differences in the dependent variable, as between a control and an experimental group, shows that the treatment did not cause the changes that occurred in the experimental group.

Impact. The effects affect whom? The impacted population is very often not the same as the targeted population--just as the actual effects of the program are often not the same as the intended effects of the program. Full consciousness of this point has not been characteristic of TSS evaluators because they tend to focus on goals in terms of population and effects. Once loosened up by sensitivity to the difference between evaluation and looking for what was intended to happen, the CND evaluator becomes hypersensitive to such phenomena as leakage, impact on siblings, parents and friends of a target student population, etc.

Durability. The long term effects turned up by the follow-up study is traditionally a strong point in good TSS studies.

Generalizability. It used to be impossible to get evaluation reports published in the only journals to which they were remotely relevant, namely, the TSS journals. The main reason given for this was that they lacked generalizability. This was frequently false, absolutely and relatively; and thus revealed how superficial an analysis of generalizability the TSS approach utilized. Today, however, the TSS evaluators are more sophisticated about generalization; they understand that evaluations can run the whole gamut from highly generalizable studies to highly particularized studies.

Statistical Significance. This is a necessary condition for the utility of an evaluation, erroneously supposed by many TSS evaluators to be a sufficient condition. For decades evaluation reports were submitted in which only statistical significance was offered as evidence of educational significance; raw gain scores would not even be mentioned. Furthermore, as recent discussions have amply illustrated, the real point about significance levels was completely misunderstood by most TSS evaluators, and used as a crutch when they were often completely inappropriate. But the calculation of statistical significance is something that TSS evaluators take in with their mother's milk, a circumstance which may be responsible for their tendency to suppose that it provides a complete diet for an adult.

Process. Process evaluation, as commonly used, is a surrogate for payoff evaluation and its acceptability depends entirely upon the strength of the correlation between the process variables and the payoff variables. In the evaluation of teachers, for example, no known correlation exists between process and payoff; and, hence, evaluation by visiting classrooms is essentially invalid. TSS evaluators have often been very sensitive to this point and have eschewed process evaluation for this excellent reason. However, it turns out that there are half-a-dozen dimensions of process evaluation which provide important justification for doing it, regardless of the connection with outcome, and these have not been particularly noticed or accepted by TSS evaluators; for example, educational processes involve interhuman interactions and, therefore, have a moral dimension. It follows that most program evaluation must investigate such questions out of respect for the considerations of justice in the classroom, that is, with respect to sex-stereotyping, inappropriate examinations, or grading standards. There are also legal, esthetic, and political constraints that must be studied. Process observation will often provide indicators of causal connections for those who realize that reporters, participant observers, and historians can make sound causal inferences from observational data of a single event. Process observation will often be a key part in determining whether the proposed or assumed description of the subject of the evaluation is correct. In addition, process observation will sometimes turn up qualitative considerations of curriculum content, currency, and comprehensiveness. Techniques of content analysis are sometimes useful and are a tool which comes from the TSS approach.

Cost. Here is one of the variables to which at least lip service is given by TSS evaluators today, though it was not to be found very frequently in their past work. The major reason for this is most important and not difficult to find. Cost has always been viewed as one of the contaminating variables that should be kept out of the ivory tower. Even economists, who were supposed to be talking about it, were frequently referring to a theoretical construct only vaguely related to it, and they rarely provided any practical skill in real cost analysis. Usually the departments of social science, from which educational researchers normally draw their role models and methodological repertoire, regard a practical cost analysis as something which should be consigned to or left in the school of business. But in the school of business one usually finds accounting and financial management, not real cost analysis. Accounting has about the same relationship to real cost analysis as contract bridge does to contract negotiations. Evaluators usually confuse cost with price; where they are sensitive to the notion of indirect costs, they usually do not realize that indirect costs include capital costs as well as maintenance costs; and if they do get that far, they do not look at opportunity costs at all; or they look at them and do not know how to handle the difference between true opportunity costs and theoretical opportunity costs; or if--by some miracle--they get that far, they ignore or do not know how to handle psychic costs. On any sane view of evaluation, cost analysis and needs assessment would have to be regarded as representing two of the three most important components, with performance measurement (testing, etc.) as the third. As long as the TSS model dominates evaluation and training of evaluators, we can expect to see cost analysis given less than due respect.

Futures. Evaluators tend to forget the future. But what is being evaluated is frequently an entity that changes with time, and the evaluations themselves are not likely to be implemented instantly. Hence, they will frequently be used to apply to and make decisions about an entity which has changed substantially. The futures checkpoint attempts to make explicit the need to project the various respects in which the evaluand will change. Extremely important examples of this are probable shifts in needs or market, political support, support by the program producer or product manufacturer, target population demographics, and increases/decreases in cost. Only once have I seen an evaluation in which futurists were explicitly brought in to handle needs assessment because it was recognized that we were dealing with what the needs of the children will be when they become adults, not just about their instant needs for survival. Needless to say, this exemplar scarcely took over the field.

Comparisons. If evaluation is to service decision-making, it must involve evaluating the options from which a decision-maker must choose. Hence, practical evaluation must be comparative evaluation. But the TSS model has never built in that insight about the nature of practical evaluation; on the contrary, its fixation on goal-based evaluation meant that it tended to assume that the criteria of merit in an evaluation were to be provided by the goals, and it was, therefore, only necessary to measure the performance of the program along those goal dimensions. Thus, evaluation could be made of a program all by itself. A second pitfall in the TSS approach that has caused perhaps more errors of evaluation design than any other, is the strong influence of control group design. In that design, the treatment/no treatment comparison is the key one. Seeing that design as an ideal--as indeed it is in many respects--people assumed that the appropriate comparison, where one should be made at all, should be between performance with the program in place and performance without it. But decision-makers do not often make that decision; they often choose between two (or more) programs to install and if they do not make that choice, then almost always they decide between two or more new programs to replace an old one. So the appropriate comparison will often be between two innovations and between those and the *status quo* program rather than the "no treatment" situation, if indeed the latter is even a possibility.

The CND view, on the other hand, treats the *Consumer Reports* kind of evaluation as one of its favorite pragmatic models and, of course, those are highly comparative evaluations. Refinement of these notions of appropriate comparisons have led to the development of the so-called "critical competitor" category, which will often include a product or program not currently in existence but created specifically for the purpose of the evaluation, as well as the inclusion of comparisons (options) the decision-maker never considered. The appropriate comparison in doing an evaluation of a promising new computer assisted instruction (CAI) program, for example, is sometimes a hard-copy program text made up by stealing the content of the program as it flashes up on the cathode ray tube screen; for that and only that will tell one whether the hardware which is, of course, a major cost item is needed. It would not be enough to compare the CAI package with an alternative hard-copy package by some other author because the payoff from the CAI may be due to brilliant choice structuring of content. Good evaluation will often require a comparison with one or more critical competitors even when no control group can be run. The notion of comparison group is, in fact, a more general notion than that of control group; and it is one of the ways in which CND may provide a better model for the social sciences than TSS because there are other areas besides evaluation where one should be thinking about comparison groups that are not control groups.

Synthesis. The time comes when one has to "put it all together." On the primitive goal-achievement model, this really is not a complicated process since it amounts to reporting on the extent to which the goals were achieved. But by the time we have developed something that looks halfway relevant to the real world, as with the preceding checkpoints, the synthesis step represents a very serious intellectual task. What methodologies does one use? This is not something that the TSS approach has much to say about; not that some models of integrating multi-scaled utility functions do not exist, but it is unclear how to apply them to the collection of checkpoints that we have above: futures, comparisons, generalizability, needs, market, etc. At this stage, synthesis is a skill that can be taught, but about which one can say very little in the way of general methodology beyond exhibiting the most primitive procedures for integrating performance, needs, and cost. The nearest intellectual skill is probably philosophical analysis or business management, neither of which is in good standing with TSS evaluators.

Apportionment. There are sometimes evaluative dimensions that go beyond synthesis. One of these which evaluators have begun to talk about lately is the apportionment problem. Suppose you know how good a number of programs are, based on evaluating them in the previously described way. You now face the necessity of distributing a limited amount of money among these programs, which are requesting, and perhaps in some sense need, a good deal more money than the amount available. What strategy should you use for the distribution or allocation or apportionment of resources? Although traditional economics is often defined as the science of the allocation of scarce resources, it is not easy to find either a set of practical procedures for doing this within economics; nor is there much evidence that TSS evaluators have addressed it in any other way.

In conclusion, let me say I have not explained fully the many complexities involved in each checkpoint. My purpose has simply been to explain briefly what a CND approach to evaluation involves, and to give test makers a check list which would assist them in designing better tests. For I firmly believe the only way to save testing from assassination, desuetude, or abuse is to look at it, work with it, and improve it in a larger framework of evaluation. Such an approach, outlined above, is one way we might begin to do this.

# EVALUATION IN A CONTEXT OF ACCOMMODATION

Lee J. Cronbach
Vida Jacks Professor of Education
Stanford University

Let me preface my remarks by stating that my paper does not fit neatly under the title of this symposium, which was adopted after I had agreed to talk about the thinking within the Stanford Evaluation Consortium. The writing team for our projected book has not yet reached the end of the road, so I emphasize that these statements are my own. Our team is agreed, however, that all talk of "models" for evaluation is troublesome and that to advocate any "best model" is to miss the significant diversity of evaluation tasks. As for the reference to "testing," we have almost nothing to say because test scores are only one of dozens of sources of information feeding into a truly instructive evaluation.

Evaluations come in many shapes and sizes. At Stanford we have concentrated on evaluation of programs. A program is a repeatable plan for rendering service. Moreover, we have thought most about evaluations of dispersed services that are intended to aid a broad policy-making community. Probably, we would apply most of the same comments, modified to allow for scale, to the evaluation of a single curriculum in a single school. But that is more Bob Stake's province than ours. We separate program evaluation to the greatest degree possible from the evaluation of specific individuals--program staff or clients. And we set to one side the continuing operations research that is carried on the guide day-to-day decisions. My colleagues and I have been thinking about evaluations appropriate for, perhaps, the Alum Rock voucher plan, bilingual education under Title VII, the work of the California Council for the Arts, the Harvard Project Physics curriculum, or an experimental effort to reduce fear of violence in schools. Perhaps it goes without saying that these have nothing in common with an experiment to compare two painkillers, or similar technological evaluations.

We set out to understand the social function of program evaluation. That is, how does one decide whether an evaluation plan is worth the effort? Time is too short to comment on the misconceptions and unsuitable practices now entrenched both in the literature and in the institutions of evaluation. I will say, though, that one of our strongest motivations in pressing for a reform of evaluation is the sense that some enormous fraction of the studies--I might go as high as 90 per cent--are not worth the effort. The talent of many keen young social scientists is going down the drain because the task of evaluation is misconceived. Let me here, however, emphasize the positive conception that is emerging, not only at Stanford but in the writings of Weiss, Rein, Cook, the "Steakhouse" team at Illinois, and elsewhere.

A program exists or is proposed in an intellectual and political context. Decisions about the plan have to be made in the light of the context. Shall the study be brief or extended? formative or summative? maximally replicable or maximally adaptive? in-house or out-house? As dichotomies, these are pseudo-issues. Generally the best kind of evaluative inquiry is a program of investigation, a mix of studies. The evaluation planner's task is to find the right mix among styles (within as well as between studies) to enlighten persons concerned with the program.

Nearly all evaluations take place in what, following Charles Lindblom, we call a context of mutual accommodation. Only very specialized contexts fit the old image of a sharp decision that a powerful and benign manager will settle rationally, just as soon as she has facts from the evaluator.

In a context of command, the manager has authority to make decisions stick. She can recognize when directives are violated and has power to punish the violators. And this requires effective agreement with the public as to the task to be done and the paramount objectives. Simon (1957) points out that only under special circumstances is a social or managerial decision "programmable." For that, there must be an agreed formula to crank out a plan for action; pour the facts into the hopper, and out comes the decision.

Most decisions are nonprogrammable. They do not even break out into deliberate, focused decisions made by specific individuals on specific dates. Institutions and communities drift into a new activity or scheme of organization, as participants adopt new views of the program, of the social problem it relates to, and of the actions that are politically viable (March and Olson, 1976).

Educational programs seek to do a great variety of good things for the community simultaneously. Which activities get priority? Which procedures are to be adopted? Such matters are worked out accommodatively, essentially by an exchange of views and political tactics, including coalition formation. Sometimes the PSC (policy-shaping community) is confined to the professional staff of an organization; reconciliation of

28

competing value systems and personal interests is nonetheless required. On the scale of something like bilingual education, the PSC includes legislators and high executives, agency officials writing guidelines, lobbying organizations, teachers who work out plans at the school level, and parents who march in to agitate for a certain emphasis they prefer. These people all have a right to get into the action, but they cannot act in accord with their own interests and philosophies if they perceive the program wrongly or incompletely.

Action comes about as society learns. The role of evaluation is to help society learn. The learning is by no means confined to the facts the evaluator reports. The evaluation can be positively miseducative if it turns into a horse-race trial of programs bearing different labels. Only a tiny fraction of the usable information an alert observer can collect is captured in final averages and significance tests.

From our point of view--manifestly inherited from Ralph Tyler--one of the evaluator's most important and continuing tasks is to learn from members of the community about their hopes and fears. How do they perceive the social problem and the current remedies? The informants to be consulted--and thereby encouraged to develop their own thinking--include political constituencies, program operators, policy makers, and suitable specialists. Questions are not confined to goals. It is necessary to trace out how these informants expect the program to achieve effects. Informants asked to think about the chain of events, from appropriation, to guidelines, to program operations, to client reception, to change in client, are led to recognize innumerable working assumptions. Moreover, as the evaluator brings the views of diverse informants before the PSC, it learns a lot about the issues--before a single "fact" has been collected.

There is a large difference between "the program"--the one an experimenter would propose to compare with some alternative--and the actual services delivered. In each site, the program is realized in a different manner. At best, then, a summary statistic can only strike an average over the specimen realizations that occurred under this year's rules and social conditions. The recurrent finding is that differences among schools within a program far overshadow differences between programs. This occurs largely because the program is carried out in so many different ways (and sometimes not at all).

The evaluator who sets out to report program experience can act only as an historian. He is studying how a program functioned at a certain time, in sites having particular characteristics. The functioning is affected by local circumstances that no one thought to mention as significant variables when the study was set up. My favorite example is Chelimsky's post hoc explanation of the remarkable variation from city to city in results of the High Impact crime prevention program: When the program was directly attached to the Mayor's office, it was productive; otherwise not. To study a program in its various realizations to see what went according to plan or hope, what did not, and why, is historical investigation. There is no way to identify many of the relevant factors in advance. There is every reason to assume that both the program and the social context will change from year to year. Indeed, if the program does not change, the evaluation must have missed large opportunities for significant discovery and influence.

The evaluator is more than an historian since he can draw on systematic methods of observation and measurement. That gives comparatively accurate information on the variables guessed in advance to be important. I do not see this systematic effort as directed wholly to the students or other clients, nor as directed exclusively to characteristics at the end of treatment. Initial characteristics and intermediate activities are candidates for systematic assessment; and characteristics of the operating institution and community had better enter into consideration. (Time has pressed me to leave out examples, but here I can tuck in a succinct one. Would any sensible person, in retrospect, evaluate the Noble Experiment of nationwide prohibition by change in per capita alcohol intake, before 1919 and after?)

Of course an evaluator cannot investigate everything. Difficult choices have to be made in selecting sites and variables. Moreover, as he proceeds, the evaluator discovers what questions are most important, theoretically and politically. Therefore, plans ought to be far more flexible than ordinary writings on evaluation design suggest.

In the context of accommodation, the community is badly served if the evaluator intends to deliver only a single technical report to the administrative agency that commissioned him. The hope of settling complex social agenda by a once-and-for-all experimental trial is mistaken intellectually as well as politically.

The public continues to think about its agenda during the time an evaluation proceeds, and whatever action follows the evaluation will not settle the matter. We are discussing social concerns that have gone on for decades; one of these concerns can become a problem at any time. For the following important sentence, postsecondary education comes to mind as an example. Even a service that an evaluation approves highly will be forced to change as public priorities change, as new kinds of clients have to be served, and as new ways of rendering the service are invented. As a matter of fact, even when an evaluator is sent out to evaluate a trial program that supposedly might be enacted into law on the basis of his report, the agenda have changed by the time he reports. Probably, the program itself has continued to change, so that his data are obsolete. And alternative program ideas that did not come under his eye will have recently entered the competition for legislative favor.

We should be looking on evaluation not only as politics continued by other means but also as social science continued by other means. In the large perspective, what matters is to comprehend the social problem and the social structure, not the service the immediate program rendered. Important also are the organizational arrangement for services and the functioning of the policy. In forming views about either the program or these larger matters, members of the PSC are influenced by information of many kinds in addition to that from the evaluator. Moreover, they must respond--if at all--not to the evaluator's words but to the clashing interpretations of his data that commentators and professional critics circulate.

The view that prevails in the community ought to be an enlightened one. It should be cognizant of the questions and answers of greatest

32

importance, and it should know which arguments rest on weakly plausible or downright suspect assumptions. Unfortunately, present machinery does not promote that kind of insight. Statements shoot through the air like skyrockets from scattered launchers--crisscrossing, illuminating the scene fitfully in different colors, never showing it whole. Evaluators are at times pressed to deliver reports that are less than the whole truth, and not completely candid about the uncertainty they properly feel (House *et al*., 1978). Given the best of opportunities, evaluators will have blind spots, and under present arrangements a critical colleague is ordinarily not able to make a public correction until the evaluation has faded from public memory.

The scientific community has a basically successful system for updating its picture of the world. Conflicting reports are compared in seminars, argued over in journals, or checked by fresh experiments. A report that is accepted gains added meaning by being imbedded in the scientific problem of which it is a part. This system of collegial review develops a consensus as to what inferences can be trusted and an equally important near-consensus about the crucial uncertainties. But the scientific community achieves mature thought by allowing itself endless time; if it takes twenty years to resolve a puzzle, so be it.

The evaluation community is asked to provide an on-line, real-time service to the polity. Scattered reports on tiny projects splintered off from social research and policy discussion often have no impact, and they risk having the wrong impact. There is no more pressing need than to invent, for the evaluation scene, institutional devices that will work like the collegial system of science, but work speedily (Moynihan, 1973, p. 279).

At the risk of offering a plan more definite than I am prepared to defend, let me suggest establishing a national panel on (say) education for competence in language. The rotating membership would include practitioners, relevant academic specialists, methodologists, and persons knowledgeable about the pertinent political-administrative scene. That kind of college, continually fed program ideas, observations from experience, and research evidence would educate itself. Given that background, it would be a proper forum for hearing new proposals or new research reports, and debating how much confidence in them is warranted. The college would be much aided if the data from any evaluation were handed round to analysts independent of the evaluation team before any report is made.

Ideally, this college would work not by issuing collective, dangerously preemptive statements, but by encouraging its members to speak their own thoughts in an instructed and responsible manner. We have had these groups in the past, but always on a short-term basis. The recent panel on the decline in SAT scores is an example. The Study Sections the National Institutes of Health use to review project proposals are continuing bodies, but they deal only with research funding, not implications. The spirit in both these kinds of groups is toughminded, collaborative, mutual education. I would hope that the continuity of scientific peer review could be grafted onto the scholarly-political mix of the typical panel looking into a social issue, in order to establish a proper audience for evaluations in any critical field.

To sum up: An evaluation of a program that is or rightfully would be politically controversial is to be judged by the quality of discussion that ensues. Simple questions about social programs are basically false questions (Moynihan, 1973). Any insight we gain about our tough social tasks will be far too complex to fit into a headline. Jim Coleman once said it (1972) in words close to these: To be useful, a social research study will have to be correct, comprehensive, credible, and _comprehended_. If, as I have said, the evaluator is an educator, that epigrammatic charge puts him where teachers have always been: attempting the impossible, but knowing the glory of trying.

## REFERENCES

House, Ernest R., Gene V. Glass, Leslie D. McLean, and Decker F. Walker. "No Simple Answer: Critique of the 'Follow Through' Evaluation," *Harvard Educational Review*, 1978, 48, 2, 128-160.

Coleman, James S. *Policy Research in the Social Sciences*. Morristown, N.J.: General Learning Press, 1972.

March, James G., and Johan P. Olsen. *Ambiguity and Choice in Organizations*. Bergen, Norway: Harald Lyche & Co. A/S, 1976.

Moynihan, Daniel P. *Coping: Essays on the Practice of Government*. New York: Random House, 1973.

Simon, Herbert A. *Models of Men: Social and Rational*. New York: Wiley 1957.

# THE VALIDITY OF DISTRICT-WIDE TEST MEANS

Robert E. Stake
Director, Center for Instructional
Research and Curriculum Evaluation
University of Illinois, Urbana

In September of 1978, at this very minute, to the best of my knowledge, there is no established evidence of validity of achievement-test means or other district-wide summary scores for indicating what might be wrong with the curriculum or for indicating how instruction might be changed to improve things. Most testing people know this and consider it only a lack of formal evidence. They may be far too optimistic. The tests appear to have little diagnostic value for district decision making even though claims are widely made that state testing programs and district assessment systems will provide a basis for improving the curriculum.

At this time we do not have the technical grounds to claim legitimately that there is diagnostic value--in norm referenced testing or in criterion referenced testing--on a school by school or district by district basis. When we look for observational evidence in the districts we find essentially none of the tests providing diagnostic information to curriculum coordinators--nor even for that matter, to most teachers who, of course, have a more focused instructional task.

But this paper is not about teachers, it is about districts.

Achievement test information available to school district people is widely used to check whether scores are as high as they have been, or as high as they are elsewhere, or as high as common sense tells them scores should be. While discussing test means, curriculum coordinators may discover problems needing investigation or solutions needing trial. Many administrators, testing specialists and researchers consider such reviews to be valuable. I think they are too. The test information itself can be useful not only politically as grounds for initiating funding changes, staffing changes, or changes in educational objectives, but the means sometimes do indicate actual change in the performance level of students.

In this presentation I am talking about group scores, not individual pupil scores. There is good evidence that for many purposes many achievement tests provide valid individual student information. The group scores are probably valid for some purposes as well, but at the time being, we do not know what the validities are. We cannot with confidence use these tests as criterion variables for evaluating the effectiveness of instruction or for indicating the quality of education a district is providing.

Many educators suspect this to be true, but as with the Emperor's new clothes . . . .

The confidence people have in using test-score means as an indicator at the school or district level is based on the professional experience of the educator using them. We do not have empirical studies that show that test means are effective indicators of school or district performance. It is not clear that such empirical studies are needed--perhaps professional experience is the best basis for accepting a performance indicator. However, districts should not imply that the validity of these summary scores has been established by the test developers or by researchers, for it has not.

What I have said so far is that the present day achievement batteries do have an experiential or common-sense validity for indicating whether or not certain district-wide performances are changing. They do not have an established validity for indicating what should be done about student performance changes. Some reviewers of district profiles look at the items or sections missed most and suppose that there is where greater instructional stress should be placed. Maybe it should. There is no established validity, however, for the test score meanings implicit in that strategy, or any other that curriculum and testing people might now recommend.

Is it that the test means have no diagnostic validity or just that we have not studied them enough? I do not know. I suspect that they do not have such a diagnostic validity.

Most educators have similar suspicions. The research of Hastings and Hotvedt showed teachers spending little time going over the results of testing. If they were interested at all, they looked to see whether scores were going up or down, and seldom looked further. Most teachers I talk to tell me they cannot expect to find information useful to them for instructing the children.

It is true that many teachers and many administrators do not understand the technical aspects of testing, but that probably is not the main reason they treat mean scores as containing little educational information. They know they are not going to find out much--at least not much that they did not already know about what the teaching ought to be.

* * * * *

In the Fall of 1977 the nation received the report[1] of the study of College Board score decline. It was greatly revealing; one of the most meaningful educational research studies of recent years. We were told "no _one_ cause could be found." That was the not-surprising but important insight of this study.

Here was a perfect opportunity for the testing researcher. The Scholastic Aptitude Test used by the College Board for applicants to college is perhaps the most carefully-built standardized test in the world. Its validity for predicting how well an examinee will perform later in college is well established. Between 1964 and 1976 national means (based annually on many thousands of examinees) had dropped slowly but steadily. Ample research money was available for studying the drop. A distinguished social servant, Willard Wirtz, chaired a panel of outstanding educators and social scientists to do the work. A staff associate, an old friend, Paul Barton, assured me it was a thoughtful, hard-working group.

Why were the mean scores going down? Were the schools doing a poorer job? Were more of the less able high school students now seeking admission to college? Were students no longer trying their best to get high test scores? How valid was the drop in test scores for evaluating American high schools?

The panel established that the tests were suitably equivalent from year to year. The drop in scores was real, but what did it mean? They found the population of student examinees had changed, especially in the 1960's, but this was not seen to be the outstanding cause of the score decline. In fact, no outstanding cause was found, just numerous contributing factors. Yes, the attitude toward testing had changed. Yes, the attitudes toward school had changed. Yes, curriculum reform had changed the aim of teaching to some degree away from what the tests emphasized, etc., etc.... What was clear was that this great aggregation of student performance data, analyzed at length by the most able of scholars, yielded no answers to evaluation questions about the quality of teaching in the American schools over those years.

---

[1] Willard Wirtz _et al_., _On Further Examination_. Report to the Advisory Panel on the Scholastic Aptitude Test Score Decline, (New York: College Entrance Examination Board, 1977).

The College Board situation centered on the analysis of a scholastic aptitude test rather than an achievement test, but I believe the results would have been no different had they studied annual scores from any standardized achievement test in any particular place. If that is the case, is it reasonable to suppose that a single testing consultant, or district research team, or committee of teachers can find causes for student performance by looking at district-wide test scores? It is my belief that standardized test-score means are determined by causes too complex and too remote for analysis. A mean is a temporary end-point for a myriad of "causes." Seldom, I think, will those means be suitable criterion scores for evaluative research studies, other than to indicate gross changes in level of performance for a limited area of school learnings.

* * * * *

For the past two years a team of us based at the University of Illinois has been studying science teaching in grades K-12 in the United States.[2] Our method primarily has been naturalistic, yielding eleven very nice case studies of district science instruction. We became interested in the role of the district curriculum-coordinator, though our attention was not usually on the district office. It was apparent to us that there was not much communication between the curriculum people and the testing people--no jurisdictional disputes or anything like that, just a lack of common territory. The curriculum coordinators we encountered did not pay much attention to test results other than to check on level of performance, overall or special subscores. The testing people sometimes were involved in microanalysis of specific lesson objectives, but almost never were concerned about major review of the curriculum.[3]

Sometimes the curriculum coordinators confided to us that they had strong reservations about an objectives-based curriculum and competency-based testing, but had to go along to protect "their turf, their subject matter" when funding or staffing decisions were being made. It was apparent to us that many administrators presumed the validity of the testing program and even subtly encouraged the testing people to overpromise what the tests could do to evaluate the curriculum.

To study the situation further, I persuaded a mature and competent graduate assistant, Mr. Norman Bowman, to find a good example of a school district where testing information was being used for curriculum evaluation and development. He first looked for consultants or trouble shooters, people who had reputations for discerning, from aggregated test data, information that was a suitable basis for modifying instruction. Some administrators told him no such people existed; some said curriculum analysis based on testing was accomplished by committee, not by an individual; and some said it was a common ability among master teachers. When he interviewed some of the nominees, he found only people who used tests as an indicator of rise or fall in performance, but not one who used tests for group diagnostic purposes.

---

[2] Robert E. Stake and John A. Easley, Jr., Case Studies in Science Education, 2 vols. (Washington, D.C.: U.S. Government Printing Office, 1978).

[3] In a separate study of district Research and Evaluation Offices, Catherine Lyon of the UCLA Center for the Study of Evaluation found testing people more linked to administrative offices than instructional offices and consequently less involved in curricular questions.

He then located himself in one suburban district where the schools were well funded, the testing program active, where a nationally recognized consultant in testing was regularly involved, and where the superintendent was both widely respected and committed to using technical resources (such as testing) to benefit instruction. Mr. Bowman observed, interviewed and studied the records for six weeks. He did not find a single instance of curricular change (or recommendation for change) directly based upon student testing. Test performances sometimes signalled it was time for change but not what changes should be made. Bowman's search, of course, was a small one, and he could have been somehow blinded to what was there. But it seems safer to me to conclude that curriculum change based on test scores is a very rare event.

Across the country today there are many state and district assessment programs based on the assumption that district personnel are ready and able to use group scores for diagnostic purposes. It should surprise some people that Mr. Bowman was not able to document how even a few such people are presently using test scores to direct the revision of curricula.

Mr. Bowman was aware that some districts are developing new curricular materials to be coordinated with test questions that will be asked in the future. Numerous large-scale curriculum projects have been so organized, Bill Gorth's Project CAM and Jerrold Zacharias' Project TORQUE come to mind.[4] But this is the development of curricula based on notions of what the students should be able to answer, not on what the students did answer. What those folks have done has little to do with building curricula from district means actually obtained.

It is difficult to estimate the validity of district means for a purpose that is not being pursued. But the question is of interest because enormous investments are being made in testing programs partly justified by claims of diagnostic validity.

* * * * *

Perhaps we should consider for a moment whether or not a mean could represent the quality of district instruction. Can we think of examples when the mean poorly indicates the quality of the whole?

> Is the mean strength of earth-filled dams an indication of the quality of work of the Army Corps of Engineers?
>
> Is the mean ripeness of tomatoes at the supermarket a good indication of their quality?

---

[4] Bill Gorth is at the University of Massachusetts. Jerrold Zacharias is at the Educational Development Corporation of Newton, Massachusetts.

> Is the mean competence of superintendents a good indication of the quality of school administration in the state?

It is apparent that means can give too little information. Even though as a statistic the arithmetic mean is sensitive to each individual case, it may not be sensitive enough for our purposes. Means also draw our attention more to the entirety of the group rather than to the circumstance of each and all individual members within the group. Should any one dam break, should the tomato any diner gets be rotten, should any one superintendent be without compassion, the whole is greatly affected. Education is very much a business of attending to the individual. Excess in one regard is not nullified by excess of the opposite kind. Efforts to evaluate education by concentration on means not only are simplistic, but may misdirect our attention as well. So we might have reservations about ever validating mean test scores as indicators of the quality of district, or school, or even classroom instruction.

* * * * *

Professor Cronbach has contributed as much as any person to the clarification of validity of educational measurements. He has stressed that validity lies in the meaning we attach to the test scores we get. A given test will have, not a single validity, but many validities, depending on the uses to be made of the scores, depending on the student population tested, depending on which test information is being used. The mean score for a district does not necessarily have the same validities that an individual test score has.

45

Many districts use an achievement test such as the Comprehensive Test of Basic Skills. It has an established validity for indicating certain achievement of certain students. In the experience of many educators the individual CTBS score is a good indicator of individual student attainment. It has occasionally been statistically validated by researchers. Never, to my knowledge, has the validity of the mean of the test been established empirically--that is, to have researchers substantiate with data that the mean indicates something about the quality of teaching or the quality of learning in that district.

The validity of a mean cannot be calculated from the validity of individual scores--even though the error of measurement of a mean can be calculated from the error of measurement of single scores. Validity is based on actual use. One cannot suppose that the meanings of the district score will be the same as those attached to the individual student score.

The greatest apprehension about test means probably occurs when they are published in the newspapers. The meaning the public often gives to a change in scores is that it indicates a general change in quality of student performance. A great number of changes occur because the group of youngsters tested changed composition or because the test changed somehow. Changes in student attainment are sometimes reflected on the test too, of course.

By and large, standardized test items draw out some of the competence the student has acquired, but only a small part. If the teachers have been preparing the children directly for the test, then the performance should reflect that learning pretty well; but it will not show how well they are acquiring all the basic skills, the general knowledge, and particularly, the complex associations we expect of human beings, young and old.

If the district's teachers have been doing what teachers traditionally have done to teach skills, knowledge and experiential associations, then the tests will test only a small portion. Most testing people believe that these limited tests still are good indicators of the general achievement of school work. They point to the high correlation between tests covering different knowledges and competencies, but they gloss over the fact that the correlations tell us whether or not the same people do best, not whether anyone does as well as they ought to. The fact of the matter is that the district achievement test mean may be a very poor indicator of how well the curriculum has been chosen, how well it has been taught, and how well it has been learned.

We just do not know. The validity of these means has not been established. They mean something, of course, but a superintendent or curriculum coordinator who changes the curriculum because of student performance means is either very experienced in matters of curriculum and instruction, or a good politician, or a gambler, and possibly all three.

Test means are products of a highly developed technology; but using them for curriculum diagnosis or program evaluation change is an act of faith, a reliance on experience and intuition, not grounded in validity studies that researchers have completed.

# PART II - RESPONSIBLE REPORTING ABOUT TESTING: SUPERINTENDENTS AND THE PRESS

48

# TESTING AND ITS AUDIENCES: CONTEXTS FOR REFORM

Frank J. Macchiarola
Chancellor
New York City Public Schools

Testing--when, how, and what to test--is the major issue in American education today. Students are now tested at almost every level of school life: when they enter kindergarden, before they graduate high school, before they enter college. Today, we not only rely on testing to determine the potential of our children, but also to measure the success or failure of our schools in developing this potential. As testing has proliferated, so has its audiences. The reading and interpreting of test scores is no longer an exercise engaged in by a few testing experts and school administrators. Teachers, parents, the general public, and students themselves want to know more about testing procedures. In particular, they want to know that the information which tests provide is reliable; and, ultimately, they want to be assured that testing serves a laudable social purpose.

The general public's interest in testing is, of course, coincident with its growing interest in education in general. This became particularly apparent in New York City in 1970. At that time, the nation's largest urban school system--with almost 1,000,000 students, 100,000 in staff, and more than 1,000 schools--participated in an experiment known as "decentralization." Under decentralization, 32 local school boards were established, with their members elected by the community. At present, these boards appoint community superintendents and control elementary and junior high school matters, subject to minimum educational standards set by the

Chancellor and to overall review of policy by the Central Board of Education. The Central Board continues to control high schools and special education schools and classes. As a result of decentralization, new policy-making centers have been formed, and because many boards are composed of parents, parent power has grown.

The phenomenon then of a system which is massive and, at the same time, local and personal has had an important effect on the kinds of testing information we need in New York City. The Central Board seeks information on the performance of New York City schools in general; local boards seek information on their schools in particular; principals need data on their students; parents want information about their children. The same information cannot satisfy all of our needs, nor can the information be presented to each of us in the same way. Certainly, the testing information we receive must be cast in terms of New York City, its unique student population and problems. And in New York City, we now realize that it cannot be the one piece of information required annually by New York Law and published faithfully in The New York Times: the reading level scores for each grade in each elementary, intermediate, and junior high school, in rank order.

A general report of reading scores is no longer satisfactory to a concerned public. The general public, as well as parents, want to be assured that tests serve some worthwhile purpose, a purpose beyond that of keeping employed those who compose, administer, and interpret tests. If such assurance is to be forthcoming, testing information must satisfy the needs of both of these groups.

First, public interaction with education is centered on one school at a time. Therefore, testing as an exercise in assessment or evaluation must focus on individual schools as the most understandable part of the school enterprise. Given this focus, the public then wants the interaction between teaching and learning, as described in test results, to be easily comprehensible. In effect, test results must be translated into general English and must concern the skills and knowledge about which the public cares.

Parents' interests are quite different from those of the general public. They want to know just what the school is doing for, or to, their child. Parents want, and are entitled to, tests that are diagnostic and test results that are easily understood. Such tests must also provide foundations for educational strategies directed by principals and carried out by teachers.

It is probably safe to say that, taken as a whole, the literature on testing does more to confuse and intimidate parents and the general public than to enlighten them. What is needed by all concerned with testing is an attitude that tests and test results which cannot be discussed in simple English sentences are probably antithetical to the public interest. Those involved in testing have a major responsibility to translate their stock-in-trade into language which can be understood by the general public. Otherwise, the Congress or some other political body will do it for them, with the strong possibility that they will do it improperly.

To implement these suggestions, the highly-centralized and, essentially, ethnocentric apparatus that is now responsible for governing and administering the testing industry must be dismantled or reformed. Urban schools cannot be properly served by tests normed on an imaginary population, conceived of in an idyllic, wooded Princeton. It cannot be served by a reporting system which locates half of the students above the norm and the other half below--a result which may make sense to the professionals involved, but indicates massive failure to parents and the general public. The failure to communicate at such a basic level should remind us of the significant lack of dialogue which exists in the testing field. Furthermore, there are some important changes taking place in our society that make the reform of testing more pressing.

One of these changes is the national move towards minimum competency testing as a prerequisite to promotion and/or graduation. Most of the pressure for such testing programs has come from the general public, and not from professional educators. The minimum competency movement provides us with another testing challenge. We must develop tests that are locally relevant and useful, and we must do so without sacrificing the general or universal applicability of the minimum competencies to be assessed. It will be a serious problem for testmakers, but one which cannot be avoided, or, as is quite often the case, ridiculed.

Another development that puts pressure on educators to reform testing is the current national discourse on federal aid to public education. That discourse now seriously doubts the efficacy of keying aid to the income and social class of public school pupils. Good public policy suggests moving to school and pupil performance as a basis for funding, eligibility and distribution. Such a move, however, cannot be predicated upon the data now available. This desirable change depends directly upon the development of instruments of assessment that are far more reliable than those that are now being used. That the Congress of the United States is currently serving as forum for this discourse adds a special note of urgency.

Similar urgency ought to characterize our professional response to the national discourse on affirmative action in schools--whether for custodians, teachers or superintendents. This discourse turns on the reliability of tests that now certify individual eligibility for a full range of jobs in the public sector. Can we not establish locally grounded criteria for school personnel that will better enable school systems to match teachers and schools, districts and superintendents? Recent judicial decisions calling for a more accurate measure of equality in the allocation of financial resources in the states and cities will be satisfied more easily if better evaluation instruments are developed.

Obviously, affirmative action in college and professional school admissions is an area of great controversy, as illustrated by the hullabaloo over the Bakke decision. Standardized testing remains a principal gatekeeper in the whole of the admissions apparatus. Schools and students both are entitled to tests that are more reliable, relevant and valid than is presently the case.

That testing is pervasive in the United States is an undeniable and perhaps irreversible aspect of our culture; that testing must remain an occasion of obfuscation and inequity is merely dependent on our collective will to reform. Testing is, after all, a great advance in the apparatus of assessment and evaluation. Let us hold the gains in education that have derived from testing. But let us not jeopardize future progress by a failure to see that if the testing industry does not move to greater self-regulation and reform, and others, necessarily less qualified, must move to do so.

In this case, self-regulation and reform really means developing and implementing a conception of testing that allows standardized measures to be locally developed (with professional help), locally administered (with professional help), and locally relevant. Such reform requires a major dialogue with the audiences for whom, after all, the testing industry performs.

# RESPONSIBLE REPORTING ABOUT TESTING: A DEPUTY SUPERINTENDENT'S PERSPECTIVE

Harry Handler
Deputy Superintendent
Los Angeles Unified School District

Testing, particularly in education, has become front-page news in the last few years. Unfortunately, such news is not always reported accurately. School superintendents may bias their reports on testing to the ideal, what should be rather than what is; whereas, the press, to create headlines, may focus on the worse aspects of testing news. What seems obvious is that testing is a vital part of American education and that testing news cannot reflect the bias or vested interest of any particular group. The press and school administrators have a shared responsibility to all sectors of the community to report not only what is presently happening in testing but also what should be happening in testing, what the ultimate goals of testing are. I have identified five points which I will briefly describe and which I believe will assist school superintendents and the press in developing a more responsible cooperative approach in reporting test results to parents and the general community.

First, the reporting of test results should be part of a total reporting process, rather than a series of independent events. To achieve this goal, an ongoing relationship should be established between the press and potential audiences to provide information and develop continuous understanding of a particular school district's testing program. The potential audiences should be given the opportunity to receive in-service training to ascertain what information tests can provide, what information we must obtain through other sources, and the relationship of the test to the instructional program. The press must also have the opportunity to discuss the instructional program and its relationship to the testing program on an ongoing basis. Continuous communication between the press and the district is necessary if the press is to understand the district's reasons for selecting certain tests, the manner in which these tests are administered and scored, and the information used to improve the instructional program. All of these contribute to the manner in which test results are reported by the press.

Certainly, the press will wish to avoid any efforts by the district to co-opt its functions in this area. But the usual press conferences, held prior to the release of press results, are an inadequate means of describing, or developing some understanding of, what most districts are attempting to do. An example of how valuable this continuous information approach can be to informing audiences about the purposes of testing programs and assisting individuals to understand testing results is to be found in the Title I parent advisory councils in my own district. Our Title I parents have a sophisticated understanding of our testing program because there has been this continuous dialogue.

Second, the reporting of results should be in "simple terms" rather than "complex language." The manner in which we report test results and the different approaches to constructing tests has further complicated the reporting of results. A brief list of different descriptors might include percentiles, stanines, raw scores, percent gain, medians, means, quartiles, standard deviations, norm referenced, domain referenced, objective and criterion referenced, etc.

We need to do a better job in distinguishing between the information that is needed for the technical analysis of test results and the information that the community is asking us to produce. I realize that many times "simple" information does not provide an adequate picture of the results, but nevertheless, our present approach is suspect.

In our district, the community has told us that they are primarily interested in the number of students who have improved, the number of students who have regressed, the number of schools that have improved, and the number of schools that have regressed. They have asked us to describe performance within the context of the following information: dropout rates, the number of students who graduate, the number of students who go on to community colleges, the number of students who go on to four year colleges and universities, the number of students who receive scholarships, the number of academic course offerings at each school, and the number of students enrolled in academic courses in each school. In other words, they are looking for school profiles and district profiles.

This approach can be very helpful if the objective is to collect data for the purpose of planning how instructional programs may be improved and not to provide data for critical school comparisons.

Third, the reporting of test results would be greatly enhanced if each district had the capability of reporting longitudinal data for cohorts as compared with data for grade levels across years. One of the most confusing aspects of our present reporting system is that we are reporting data for third graders this year and comparing this information with that of third graders last year, instead of having an ongoing system where we can follow the performance of students throughout their experiences in our schools. I would imagine small districts have already developed a means of reporting results across grade levels rather than comparing results for different youngsters in the same grades at different times. Unfortunately, we have not reached this point.

Comparison of the same grade levels across years further tends to confuse the public and to place district personnel in a position of hypothesizing the variables contributing to differences, rather than being able to relate changes in the instructional program that have been implemented to meet the needs of changing student populations. The validity of comparisons of students in the same grade levels during different years, especially when tests are changed en route, severely limits the district's ability to interpret results to the press and the community.

Fourth, the reporting of test results could be improved if a greater degree of stability of reporting procedures, over time, were to be included. When districts or state departments make changes in the testing program--for example, changing from norm referenced testing to a matrix sampling approach--we surely confuse the interested audiences. Although the professional educator may wish to continuously stress the importance of determining the extent to which individual students are improving, rather than comparing students, classes, grades, or schools, parents and the community will continue to ask for comparative data. When those responsible for the legislation or implementation of testing programs, continuously change the tests, the procedures used in developing the tests, or the manner in which the results are reported, the public once again loses confidence in the integrity of the schools. In this area, whatever we as a profession finally decide to do, I would suggest that we have some stability so that we reduce the confusion in the community, rather than lose credibility under the cloud of "intentionally changing each year to avoid telling it as it is."

We will not have responsible reporting until all of the "players" are willing to make some responsible decisions about which base line data we will use, which results will be reported and for what purposes, and provide districts with a reasonable period of stability, at least three to five years, during which we can bring some semblance of order to our reporting system.

Fifth, responsible reporting requires planning in advance of the administration, scoring, interpretation, and reporting of test results. Our ability to report information will improve when we devote greater attention to advanced planning related to how information from testing programs will be used to improve the instructional program. This planning should take place even before we know the results. Unfortunately, it has been my impression that many times the "what to do with the information" is, *post hoc*, the sharing of the information with the press. This places districts in a situation in which a major portion of the press conference is devoted toward "explaining away" when the results are poor. I believe that we should be prepared to describe the results in as honest and forthright a manner as possible and, probably more important, to anticipate the results and be able to describe to the public the next steps to be taken.

If any of these points have any value, I suggest that they be shared with the audiences of concern as soon as possible. We should begin to negotiate with our interested audiences what we "can do" and what we "are unable to do." Congressman Albert H. Quie has stated it well in his article, "An Elected Official Looks at Mandated Assessments":

> "....Assessment and the attendant accountability are a two-part phenomenon in which one party demands--or makes known his wish for--accountability, and the other party responds out of a sense of 'oughtness.' Accountability does not readily come if it is not asked for or mandated. So both parties will have to become educated--the one as to the ways and techniques of being accountable, and the other as to what can be expected in the way of accountability, the forms of mandating it, and the information that is required prior to asking for it. This is the human exchange in which we are now involved."

It is possible that if school superintendents and the press further enter into this "human exchange" that both will be able to markedly improve the reporting about testing to their constituents.

# RESPONSIBLE REPORTING ABOUT TESTING: AN EDUCATION CORRESPONDENT'S PERSPECTIVE

Ronald D. Moskowitz
Education Correspondent
San Francisco Chronicle

As a reporter, my main task is to report the news. And news, as we know it, by its very definition, is something controversial. The reason there has been so much in the news lately about testing is because testing--from IQ to achievement testing, and from the elementary to the post-graduate level--is controversial today and is getting more so by the minute.

Look around you. The first thing that happens to most children when they enter a school system--public or private--is that someone gives them a test, or a whole battery of tests. How they do on these first tests will be crucial and critical to how they are thought of for the rest of their lives. They will be processed, stamped and labeled because of these tests. And whether they are Grade A or Grade C merchandise by the time they reach college admission age depends in great part how they did on that first IQ or reading readiness test a dozen years earlier.

If you think this an overstatement, then let me remind you of an experiment in South San Francisco a few years ago in which a group of school children were tested at the end of their first year in school by a Harvard researcher. Well, a funny thing happened to these children on their way to the second grade. The researcher, when the scores were in, altered the records so that those who had done poorly were labeled "gifted" on their records and those who did well were listed as "failures." When they entered the second grade, their teachers looked over the cumulative or "cume" folders, as they are called, for clues as to what kinds of students they would have in class. And guess what happened that year. The ones who were mislabeled "gifted" did extraordinarily well in the second grade, without exception. And the ones who were labeled as "failures" all did poorly.

The experiment was designed to see whether a teacher's expectations of a child have any effect on how well the child actually does in a classroom--in other words, whether the label pinned on a child at an early age is seen by the teacher as a predictor of how well that child will do in the future. The experiment showed that the scores children--very young children--make on a test can pin a label on them that will follow them the rest of their lives. The label would become a self-fulfilling prophecy.

The results of that experiment cannot help but make the people out there who pay the taxes that support the schools wonder how many children are also labeled as poor students simply because they could not answer standardized test questions that tested the child's culture or economic status rather than his ability to learn. And it cannot help but make minority parents wonder how many children are labeled as "poor students" by their teachers because of their accent or the color of their skin. Would any of you here deny that such labeling goes on? Or that teachers who label because of skin color or accent defend their actions by citing the children's scores on standardized tests?

In a June interview in The San Francisco Chronicle, Samuel C. Kohs, who helped develop the modern IQ test, known as the Kohs block test, some sixty years ago, said that he has been dismayed and angry ever since at how this test is used in the schools. Dr. Kohs's main objection, apart from the cultural discrimination inherent in most IQ tests, is that IQ tests are used to assign numbers to a child's worth, based on a very limited measure of the whole person. He called this "outrageous."

He said that IQ tests, and the numbers attached to them, are sometimes interpreted by teachers, the educational establishment and by children themselves as measures of overall intellectual worth. "Our tests focus on such narrow things, the ability to acquire knowledge," Dr. Kohs commented. "What of human feelings? What of social abilities? Are not these also important?"

Modern schools, he added, are designed "to teach all children in the same way, so that some of them will some day get Ph.D's and the IQ test tests their ability to do this, to acquire cognitive knowledge. It does nothing to show us how to develop all of a person's potentialities. A measure of 70 to 79 on the scale is called borderline defective. This is terrible."

Let me remind you that these are not my criticisms but those of an 88-year old scholar whose block test is still widely used today as an integral part of the Wechsler series of IQ tests. Kohs said that what is lost in modern school systems is "a sense of the uniqueness of each individual. The curriculum that 30 or 40 children are compelled to do together--instructing the mind of each child regardless of differences--is

63

something that is against human nature," he argues. Most modern educational psychologists, Dr. Kohs added, recognize that the IQ is just one of many measures of a person. Yet all too often, IQ tests, as actually applied in local school districts, "are used far beyond their inherent value."

It was one of these variant uses that caused U. S. District Judge Robert F. Peckham four years ago to prohibit California public schools from using group IQ tests to determine whether a child should be put in classes for the educable mentally retarded. Because of the tests, administered in large groups to students whom teachers found to be problems in their classrooms, large numbers of black students were being shunted into classes for the retarded. While only ten per cent of the state's school children are black, blacks make up 25 per cent of the classes for the retarded, the lawyers for the blacks argued. Statistically, they said, there should have been no larger a proportion of black students in the classes for the retarded than there are in the school districts of the state as a whole.

Once that battle was won, the attorneys filed new motions asking the judge to ban the tests permanently from the public schools of California. Recently, that issue was heard during five months of testimony before Judge Peckham, and his decision in that case is due some time before the end of the year.

The courts are not the only institutions being asked to intervene between those who make and sell testing instruments and those who take the tests--the consumers. Last October, at a meeting of the College Entrance

Examination Board in San Francisco, Joseph A. Califano Jr., Secretary of Health, Education and Welfare, announced the first federal government investigation into the fairness of standardized achievement tests since they first became popular during World War I.

The federal government is interested for the same reasons that it is interested in faulty automobiles or bad medical care or any of the other thousand-and-one items that affect the quality of life in America today. Secretary Califano is interested because he knows that one bad label can do as much harm to a child's life as a faulty gas tank on a Pinto that explodes on impact. Either way the child gets burned. He is concerned because he feels that standardized tests may be becoming the Laetrile of the educational world--the cure for the cancer of ignorance that in reality cures nothing but, in fact, causes serious side effects of its own.

The studies he announced will be undertaken by the government's National Institute of Education and two of the nation's most prestigious private groups: the National Academy of Sciences and the National Academy of Education. The studies are designed to determine what such tests measure, what their strengths and weaknesses are, whether built-in cultural biases cause some students to fail them consistently, and how they are used throughout the nation.

At the time of his announcement, he noted, "the striking lack of minority participation" in graduate and professional schools. He also attacked the heavy reliance most colleges and universities place on such test scores in their admissions procedures. Such tests, he said, are "tools, not magic wands." He did state, however, that testing for basic competency "can serve important purposes in our educational system." He warned, however, that "basic competency testing will be effective only if we stress, along with its benefits, the critical limitations and dangers of testing."

Secretary Califano emphasized that we must continually seek new ways of measuring human potential--"ways that reach beyond the traditional yardsticks." I think he echoes the chief complaints about standardized tests today--both IQ and achievement tests. In general, the complaints are not about the tests themselves, which almost anyone in the testing or educational or psychological fields will readily admit do have some inherent value. The chief complaint seems to be that too many incompetent people use the various tests far beyond their inherent value, and that is where the main trouble seems to lie.

Dr. Kohs, as I stated earlier, is distressed by the use of IQ tests today. But he has no quarrel with the measurement itself. "If you want to avoid measurement," he said, "you must take all of the adjectives from the dictionary."

What he and Dr. Califano are looking for and hoping for is a much better test than anything yet developed--what Dr. Kohs called "a test that measures true brain power, the true capacity of the human mind." "The IQ test does not truly measure intelligence," he reminds us. "It measures something, but we must get much, much better tests that measure the whole person." I would imagine that every one of us in this room would agree with that statement.

The other thing that Dr. Kohs and Secretary Califano and the citizens who brought that suit now pending in Judge Peckham's court are all looking for is some guidelines that will help the overzealous misusers of such tests to mend their ways. The studies now being made in Secretary Califano's behalf by the government and private groups may well provide these guidelines.

Whether the education industry--and it is an "industry"--will voluntarily take steps of its own to solve some of the myriad problems that standardized tests have caused in this country, or whether the industry and the companies who market these tests will force the courts and the federal government to provide guidelines of their own are unknown at this time.

I think it is time that the people who make and use standardized tests in this country realize that there are some real problems involved in testing, problems that affect thousands of people's lives daily and that someone is going to take measures to correct the inequities that exist today largely because of these tests.

These corrections can either come voluntarily from the people who profit from the tests, or they can come from the courts, the federal government, or both. At this time, before Judge Peckham rules and before Secretary Califano's task force reports, there is still time for some voluntary action.

Ironically, despite all of the problems, the American people--especially the white, middle, and upper middle-class American, for whom most of these tests were designed--have great faith in standardized test results. Witness the Bakke case, and the furor that the California Supreme Court's decision caused in that case. And remember the national dialogue that developed over testing as the appeal was pending before the U. S. Supreme Court.

Americans--be they rich or poor, white or black--are looking for some fair, objective and equitable standards by which they can judge who should get into the nation's schools and colleges, into graduate and professional schools, and who, on the other hand, should and must be relegated to classes for the mentally retarded.

The problem, as I see it, is that we have no such "magic wand" at the moment and are unlikely to have one in the form of a standardized test--or a battery of standardized tests--in the near future. Therefore, it is incumbent on everyone in the testing business to come up with some guidelines on the use and abuses of standardized tests in the very near future.

For too long, the education industry and the testing manufacturers have put themselves in the position of the auto industry, telling us we may sell the cars but we cannot control how they are driven. Well, tests, like cars, can be deadly weapons and they must be controlled somehow.

If you and the industry do not regulate the uses of such instruments, those who represent your consumers--the federal government and the courts--will surely do it for you. And we in the media will continue writing about the controversy because that is our job.

# PART III - MINIMUM COMPETENCY TESTING:  PROBLEMS AHEAD

# MINIMUM COMPETENCY TESTING:
# A NATIONAL PERSPECTIVE OF PROBLEMS AHEAD

Chris Pipho
Associate Director
Department of Research
Education Commission of the States

What are the problems ahead for state minimum competency testing programs? That depends on who you are and what you see as the problems and the solution. If you are a member of the "general public," the problem is a declining confidence in the public schools' ability to turn out students acceptably trained in the basic skills. The tenth annual Gallup poll of public attitudes toward the public schools revealed that between 60 and 70% of the adults polled thought children should be promoted from grade to grade only if they can pass examinations. A random sample telephone poll conducted by CBS for the three-hour television program "Is Anyone Out There Learning?" showed that over 80% of the more than 1,000 adults questioned wanted a test or a standard required for high school graduation.

The public's "solution" to this problem is a minimum competency testing program. As of mid-September 1978, 36 states have taken either legislative or state board action to mandate that schools on a statewide basis identify the minimum basic or life skills that students should attain as a result of moving through elementary and secondary schools. Setting minimum standards for grade promotion or high school graduation has become one of the hallmarks of state action. Not all states, however, mandate a statewide test. Generally, state activity falls into three categories:

1) a state standard and a state test, 2) a state standard and a local test, 3) a local standard and a local test.

If you are an educator--elementary or secondary teacher, administrator, superintendent of schools or college professor training teachers--the problem is the mandated minimum competency testing program. In contrast to the general public, educators want either to repeal the laws or modify the programs mandating testing, give more authority to teachers or school coalitions or ban all testing completely. In general, the public's concern over the declining quality of student learning and teaching is almost routinely dismissed by educators who insist there is no problem. Perhaps then, competency standards movement is neither the declining quality of learning and teaching nor defining which types of minimum competency programs are best, but the lack of communication between educators and the public.

Let me explore with you briefly how various members of the public and educators see the future of minimum competency testing.

First, state legislators:

Interest in minimum competency standards by state legislators probably peaked in 1978. Thirty-six states have now taken some form of action. Not many of the remaining states will choose to take action and not many legislators in these states have an overriding interest in the issue. Exceptions to this generalization are in Indiana, where Representative Jonz is currently working on an interim study with legislation likely to be introduced in 1979. Other states in which legislative interest remains high are Washington, Minnesota, and New Jersey where a bill is still pending.

It is interesting to note that New Jersey already has competency legislation enacted. The pending bill would extend the minimum competency program standards to high school graduation. This extended interest in New Jersey suggests that the minimum competency issue has both a first order (constituency) parameter and a second order (political) parameter. For the most part, the first order pressures have already resulted in legislation being introduced and passed in various states in 1976, 1977, and 1978. The second order pressures, mostly political, started to surface in 1978 and appear to be a growing force in the introduction of minimum competency standard legislation. This form of pressure was most visible in Kansas during the 1978 session. There the state board of education mandated a task force and study of the competency testing movement and took action early in 1978. It was open knowledge that legislators were unhappy with the slow moving state board and routinely announced that unless the state board and the state department of education took action, legislation would be enacted. Late in the 1978 session, during a one-day veto override session, a much-compromised bill was enacted calling for a two-year pilot testing program in minimum competency testing.

This competition between legislators and state boards of education is present in other states. While competition over the control of elementary and secondary education between education committee chairmen and state board members or chief state school officers has always been around, it has now surfaced and become more politically visible on the minimum competency testing issue. In Illinois, legislative interest remained high during the

1978 session. The state department of education, in cooperation with local districts, tried to solve the problem. A compromise was a law enacted during the summer of 1978 which gives the state department of education two years to create a minimum competency testing program. The legislation included no specifics in the mandate except that a two-year optional period was established.

Looking ahead, legislative interest appears to be moving into other areas related to student competency testing. During the 1978 sessions, several states introduced bills calling for basic skill or minimum competency tests for teachers as a part of the recertification process. None of these bills was enacted, but two states, Georgia and Florida, are currently implementing programs of basic skill or minimum competency tests for the precertification process. It is expected that legislative interest in teacher competency will continue and probably increase during the 1979 legislative sessions.

Legislative interest at the federal level, of course, has not declined. There, Congressman Mottl from Ohio modified his position throughout 1977 and 1978, and at this time still includes in the Elementary and Secondary Act, extension HR 15, a call for the support of local school districts and state departments of education in writing their own voluntary minimum competency testing standards. Many observers of the federal scene feel that it is only a matter of time until the federal government moves more directly into supporting national standards in elementary and secondary education. At the annual conference conducted by the National Assessment of Educational Progress in June of 1978 in Boulder, David Shoemaker told NAEP conferees that in his opinion--and not from the department's view--"while it used to be valid thinking for the federal government not to want to tell the states what to do, 'the times have changed drastically.'" He asserted that since many students change schools, and even their states of residence, "it is no longer sound to say that the federal government should not set standards in education," standards which would apply across state lines. His own personal prediction is that during the next five years, there will be federal standards set to respond to these shifts in student population. "How can the federal government carry out the thrust of sections of HR 15 without setting standards?" he asked, in outlining his personal opinions. "I don't see how there can be anything other than some kind of statement of federal standards for basic skills."

Second, state board of education members:

Control over elementary and secondary education is a key factor in a number of states. The budget prepared by state boards, and the control they exert over existing reading, writing and arithmetic programs and the curriculum in general is fairly well established in most state statutes. Many state boards of education can make changes in these areas without new legislation. That state boards can control certain aspects of education without legislation probably contributes much to the competition between legislators and state board of education members. Careful study and a slow review of the problems and issues characterizes state board action. When state boards do act, they usually move into the area of life skills, survival skills, and functional literacy. In contrast, state legislators usually move quickly; and when they take action, it usually involves a mandated paper and pencil reading, writing, and arithmetic test.

In many states, state boards also have the unique responsibility of implementing legislative mandates. This, of course, gives these people an opportunity to interpret and work with educators on legislative mandates. In 1978, the state of Massachusetts completed a year and a half study of the minimum competency movement, a study which involved three large task forces of teachers and educators, and which resulted in a program calling for a cooperative state and local school district basic skill improvement program. Although this program does not call for any additional requirements for high school graduation or grade promotion, it does use a test to help identify students who need assistance both at the elementary and secondary level. The strength of the state board study in Massachusetts was the involvement of all education groups in the state. As a result, support from statewide teacher associations, both NEA AND AFT, and school boards is evident. In general, state board action tends to involve large numbers of people. In six states, Maine, North Carolina, Alabama, Kentucky, New Mexico, and South Carolina, formal commission or task forces were used to help select the test and set the standards in minimum competency testing. Looking ahead, state board members are more concerned with implementation problems than they are with new mandates for tests or standards in minimum competency. Up until this point, they have demonstrated little interest in the teacher competency issue and seem to be more interested in addressing the communication gap between the public and educators.

Third, educators--teachers, principals, superintendents:

In general, educators are skeptical of the minimum competency testing mandates that exist in their states. In states where the mandate has not yet been handed down they are even more skeptical. The National Education Association has come out in opposition to the minimum competency movement. They feel that state mandated standards should set no more than broad general curricular guidelines and should not be based on student achievement. They believe that the use of equivalency tests in lieu of academic preparation and school attendance for meeting requirements for graduation should be strongly resisted and that standardized tests, used as a criterion for high school graduation, should be eliminated. In Florida, this teacher concern reached a peak when the National Education Association and the Florida Teaching Profession/NEA commissioned an independent panel to make a study of the Florida functional literacy test. The report that was issued, entitled The Florida Accountability Program: An Evaluation of its Educational Soundness and Implementation, was released in the summer of 1978. The committee, chaired by Ralph W. Tyler, using two other test experts and two Florida teachers, was for the most part, unhappy with the manner in which the program had been carried out. I am sure that Ken Loewe will fill us in in greater detail on this report.

While educator concern is especially pointed in Florida and in other states where the implementation process is now underway, it is interesting to note that in local districts that have been implementing minimum competency testing programs for one to seventeen years, the issue has not been especially heated. I have not been able to find any local school district in which the issue has been the basis of a negotiated agreement, a strike or organized teacher resistance. Teachers in the Westside School District in Omaha, Nebraska, in Gary, Indiana, in Denver, Colorado, and in other districts generally take the tests passively or state that such test motivate the middle range of students while leaving students in the upper and lower end of the achievement scale untouched. These teachers seem to say that testing gives them a motivational hold over students that the report card no longer has. The future for educators at best remains cloudy. It is quite likely that the NEA will continue its stand in opposition to minimum competency testing and could very likely commission other statewide studies similar to the one conducted in Florida, particularly if the Florida study appears to have had some success in turning the minimum competency program around or modifying its implementation effort. This group through its professional organizations will continue its pressure against a national test or national standard.

Fourth, postsecondary teacher preparation:

Involvement in minimum competency programs for elementary and secondary students by teacher training institutions has not been particularly evident. As implementation begins in many states, interest and involvement will no doubt increase. In testing and evaluation departments, this, however, is usually not the case. Gene Glass's extensive writing, speaking out on the subject of setting standards, H. D. Hoover's concern over the use of the Iowa basic skills test in Iowa, and others have given very specific assistance to legislators and state policy makers in this area. The future for teacher training in postsecondary institutions will probably include many workshops and summer institutes for school district people, involved in the implementation of minimum competency programs and some introduction to the movement and its problems to undergraduate students. Doctoral dissertations concerned with minimum competency are beginning to appear, with approximately 15 to 20 such studies currently under way or already completed in the past year. Postsecondary institutions appear to be following, not leading, in this movement.

Fifth, the public:

The polls appear to be clear: the public does want tougher standards for grade promotion or high school graduation. Perhaps it is safe to say that the public always gets the kind of schools it wants. Throughout the history of American education, public opinion has been a strong force in the creation of the American dream: educational equality for all. If educators often feel that parents view student problems very simplistically, it is probably fair to say that the parents' viewpoint should not be ignored. Parents have the responsibility for their children. They spend large sums of money for corrective shoes, braces, eyeglasses and any other aid to assure proper physical growth and safety, and it is difficult to explain to parents that they cannot have the same control over the development of their children's basic skills attainment. Parents may, indeed, view the issue simplistically, but a survey conducted by the Policy Studies in Education for a suburban Wisconsin school district found that the public does not expect the schools to do everything. They made a distinct separation between those things that students should know and those things that the schools should teach. In the survey, the public was asked how they would spend a thousand dollars of school aid in fourteen different areas:

78

nearly 1,000 residents said that they would spend $400 on basic skills, $115 on career education, and $25 on the fine arts. If this survey and other polls conducted by Gallup and CBS are correct, public concern and interest in the basic skills is high and will undoubtedly remain high.

Summary:

What lies ahead for the minimum competency testing movement? Jeffrey H. Fletcher and Gary D. Woodell, teachers in a junior/senior high school program in Milford, Ohio, in a report entitled Minimum Competency Testing: Today and Tomorrow, make these predictions: "If the minimum competency testing programs do indeed progress as forecast, certain developments in the schools are highly probable. First, the focus of what is taught may narrow to embrace only those skills which are tested by the minimum competency testing program. One immediate result of this narrowing of focus would be the loss of status of other nontested programs such as art, music, industrial arts, physical education and similar programs. This loss in status could take two equally devastating forms. The first, funding for the resources and personnel could be cut back or maintained at the lowest possible legal level (assuming that these programs will still be mandated by the states to be part of the educational program). Second, these nontested programs will probably lose status in the eyes of students, parents and educators as they are deemed not important enough to be tested. If one's graduation from high school depends on one's ability to read, write, and compute (as indicated by basics being the choice for the testing program) one will emphasize that in the educational program." They also go on in their prediction for 1983 to say that there will be "an emphasis on quantitative evaluation at the expense of qualitative evaluation; the resultant quantifying of previously unquantified subjects, such as social studies and composition; and also the teaching for the test with the result of a lowering of expectations for all students; and an overbalance of attention, if not funds, for remedial or compensatory programs at the expense of growth in 'regular' and gifted students." It is possible that this prediction for 1983 by two practicing teachers may in some ways reflect the concern that teachers in general have over this whole movement. I believe state and local school districts will, in the next one or two years, experience a number of pressures for changing the minimum competency testing movement. Some of these are likely to be:

1) A legal test of minimum competency testing programs may surface in 1979 and continue from that point on. The primary focus, of course, will be on discrimination against minorities. Court tests will probably force educators to look carefully at school programs as they relate to competency tests. The Florida situation will bear watching. Another potential area for court litigation will be from special education and handicapped students and the use of the individual education program or I.E.P. under federal law 93-142. Concerned individuals in this arena are predicting that a life skills or survival skills mandate that does not include an interpretation of survival or life skills for special education or handicapped students will be subject to court tests.

2) Dissatisfaction with state implementation efforts will continue and intensify where states have difficulty in selecting tests and standards and in implementing them in a wide variety of school districts. It is

already evident that in Florida (with a state standard and a state test) educators and others are looking at the possibility of obtaining more local control over the state testing program. On the other hand, in states such as Oregon, where local control is a key element in the state mandate, the call will be for more state assistance. It is probably safe to predict that states will, in some form or another, succumb to the "grass is greener on the other side of the fence" syndrome.

3) The trend to change minimum competency from a single test cutoff standard for grade promotion and high school graduation to a remedial identification program will continue. This will have ramifications on budgeted state dollars for compensatory education and remediation programs administrered with state funds. Making the competency testing program an "early warning" system will continue to attract more state policy interest.

4) Curriculum planners and textbook writers are already moving to develop classroom materials to support the statewide mandates. The end result will be more of a "skills continuum" approach, especially in math. It is quite likely that many of these materials will be nothing but skill workbooks with hard covers. The name of the game will be sales and speed in getting the new product to the market place.

5) Students labelled "incompetent" will be another problem. This remains a far bigger problem for those not working with a program than for those local districts where diplomas have been denied. I predict that the public is willing to stand for more nongraduation and nongrade promotion than educators think they will tolerate. Florida's plan to allow students to take the functional literacy test and receive a high school diploma at

any time should be watched.

6) School entry and minimum competency testing may be a growing area of interest. The new law passed in South Carolina this year calls for a school entry minimum competency test to alert parents of possible deficiencies.

7) Better follow-up studies of the students failing competency tests is needed. Some concern over dropouts has already surfaced. More information on the relationships between classroom teacher and tests, and between national norm-referenced tests and tests of minimum competency is needed.

8) Less formal solutions to the whole issue need to be explored. In this category, I would put parent-administered short tests of minimum competency and more and better communication between educators and parents. It might be possible to answer the call for more accountability without using tests.

9) There will be more interest in competency testing of teachers. I think it is highly probable that at least one state will enact legislation in 1979 calling for a competency test of basic skills to be administered before teachers are recertified.

10) Other concerns, Proposition 13, vouchers, declining enrollments, etc., will begin to take the public spotlight as more articles in the media indicate that test scores are on the rise again.

In conclusion, my thoughts on the subject of minimum competence testing are the same as those of the Commission of the States as expressed in its 1978 Policy Recommendations:

82

ECS supports state efforts to improve educational achievement and to increase the educational attainment of underachieving students to levels that will enable them to participate effectively in society, including obtaining and retaining a job. Setting standards for graduation, grade-to-grade promotion or placement in programs and specifying methods for determining when those standards have been met are common practices in education. Some standards are set by the state, some by the local school district or school. Teacher-made tests, department-wide examinations, standardized tests, minimum competency tests, teacher judgment, demonstration of mastery by student performance are some of the methods used to determine when pupils have met the standards. Use of minimum standards and minimum competency tests to determine when the standards have been met is one option. There is not, however, a body of knowledge or research data to support a conclusion that this method is better than others. Therefore, ECS advocates that states and local districts adopt a comprehensive plan for setting standards and establishing criteria for determining when those standards have been met. Such a plan should avoid the selection of standards that are too narrow and the reliance on a single assessment technique.[1]

---

[1] ECS Policy - 1978 Recommendations (Denver: Education Commission of the States, 1975), p.28.

# COMPETENCY TESTING: THE FLORIDA EXPERIENCE

Ken Loewe
Director of Test Development
Florida Department of Education

The Florida educational system has been, and is, deeply involved in testing. Indeed, Florida was one of the earliest states to pass accountability legislation. Our early testing programs were conducted on a state representative sample and, therefore, we issued state level reports. These state level reports were a sort of general report card on the composite Florida public school system for specified grade levels. As a general report, it had a general audience and was greeted with general indifference.

In an attempt to counteract this indifference, we began reporting at four additional levels. We reported at the district level, at the school level, the classroom level (if the teacher wished), and the student level. This reporting avoided the "they are not my scores" syndrome and helped place the focus on academic problems. But the testing and subsequent reports did not have any teeth or mandated consequences. After the initial shock wore off, local education agencies (LEA's) were able to rationalize or ignore the reports, and consequently there was little change, reform, or action based on needs identified through the state testing program.

To remedy this indifference, the Accountability Act of 1976 was passed requiring each school district to develop a pupil promotion plan. The plan must include mastery of the basic skills by each student before promotion from grades 3, 5, 8, and 11. Basic Skills are defined as reading, writing, and arithmetic, and mastery is determined by a state test.

One of the best known aspects of the law is the requirement that students pass a state examination in functional literacy before they are eligible to graduate from high school. If a student does not pass after repeated attempts, he or she is not eligible for a high school diploma. This requirement applies to all students in regular high school programs.

The Accountability Act of 1976 took effect for the basic skills in the 1977-78 school year, and the functional literacy requirement will apply to the graduating class of the 1978-79 school year.

One of the less well known aspects of the Florida program is the remediation effort that has been launched. In the last two years, 36.5 million dollars has been made available to school districts to assist them in providing remediation to students who have not mastered the basic or functional literacy skills. The districts are permitted to use these funds only to provide direct remedial instruction.

My experience with the Florida program, its problems and the problems which seem likely to surface soon enables me to discuss the broader aspect of minimum competency testing and its future problems. Indeed, many of the problems faced in Florida's program are likely to surface in other applications of minimum competency testing.

I find it convenient to classify the problems in five areas. These areas are : (1) legal issues, (2) philosophical issues, (3) political issues, (4) measurement issues, and (5) educational issues. While these issues are not independent of each other and do overlap, I find this a convenient organizational pattern. I am dealing with issues that have been raised in Florida; not all the issues are problems that may arise in other states.

## Legal Issues

I will first discuss some legal problems, issues, and concerns of the Florida program. Some people do not like Florida's competency testing program, and legal challenge is one approach being used in an attempt to stop the program. Our first challenge came from a single citizen on behalf of his son. The father was concerned because the testing program identified his son as a non-master in two specific areas dealing with punctuation.

Initially, this parent received no satisfaction from the local school board and pressed his case on the state by attacking the legal technicality regarding authorization to set the cut or passing score on the test. A state administrative hearing officer ruled that the Commissioner of Education was not authorized to set the cut score, as had been done. This ruling is being appealed in civil court. Meanwhile, the legislature clearly specified that the State Board of Education was to have the authority to set such rules.

At this point, the parent was joined by a legal team representing the National Association for the Advancement of Colored People (NAACP) and the Southern Christian Leadership Conference. The main line of attack by the plaintiffs concerns the scoring system and specifically the cut score. This is not surprising because the level of the cut score is the single most visible factor determining passing or failure on the test. At present, attorneys are still asking and answering questions of each other. This will be followed by more public hearings and a high probability that a decision in either direction will be appealed.

Minimum competency testing, which may result in retention at a grade level or in failure to graduate from high school, is a ripe area for legal challenge. The process is slow, tedious, and expensive in terms of legal counsel and staff time, but probably an inevitable consequence of instituting the kind of program we have in Florida. We expect that legal issues will be commanding our attention for some time to come.

## Philosophical Issues

The legal issues and challenges we are experiencing may really be based on philosophical issues. Does the Department of Education or the legislature have the right to deny a diploma to any student, or is it the right of every student to get a diploma? Does high school graduation serve only a social function, or is it entirely a credentialing event? Do schools have the right to retain students in a grade for lack of competence as measured by tests, or is promotion a right of all students? Is there a minimum competency that society can insist students achieve before they join the job market?

If we can get past this philosophical issue and agree that society has the right to expect performance as a precondition to promotion or graduation, we then need to determine what areas should be tested and at what levels of performance. What are the minimum skills upon which we will judge whether a student passes a competency test? Should we test for survival skills? Independent of any further measurement, we can predict that Florida students would probably do very well on a survival test. We know that a very large percent of Florida students do survive in school and in the first years after school if we define survival as "not expiring." We probably do not want only a "survival" test.

In Florida we determine minimum competency through criterion referenced tests of basic skills. But how do we define what skills are the basic skills? What are they basic to? For whom are they basic? With the Florida legislature's interest in basic skills came a flurry of interest from a variety of curriculum groups who wanted to get their special area aboard the "basic skills" bandwagon. Curriculum leaders in physical education, industrial arts, and driver education wanted their subject to be included as one of these basic skills. Lapel pins sprouted with the proclamation "Art Is Basic" and it certainly is, at least to someone. How are we to sort through these various claims? How are we to identify only the truly basic or minimum competencies?

We live in a pluralistic society, built and maintained by the application of a broad range of skills. Defining the basic skills could have resulted in an endless debate in Florida if the program had not been solved by the legislature. In the present Accountability Act, the Florida legislature has defined basic skills as reading, writing and arithmetic. The philosophic issue of what skills are basic has thus been answered in Florida through the legislative process.

The definition of basic skills was further refined by involving educators throughout the State in deciding which specific reading, writing, and math skills were basic. The educational community was given a list of skills and asked to indicate whether each listed skill was: (1) basic, (2) placed at an appropriate grade level, and (3) clearly stated. Through an iterative process we developed a list of minimum basic skills that is supported by at least the majority of Florida's educators.

This was Florida's approach to the philosophical issues. In beginning a minimum competency testing program, the same philosophical questions must be answered. Those questions are: (1) Are there minimum skills to be expected of all students? (2) How are these skills to be defined? (3) Who should define them?

## Political Issues

Politics and political issues are highly relevant to minimum competency testing. Most of the large scale competency programs that have recently appeared are the result of legislative action. This involvement of law makers with education can result in some unique problems.

We are about to face some difficult political realities in Florida. We tested about 110,000 eleventh graders in October, 1977 on the Functional Literacy Test. About 37% or 40,000 students did not pass the test. Can our school administrators, school boards, or legislators resist the public resentment that will be aroused if 40,000 seniors are not ready to graduate in May of 1979? The 40,000 students will be given two more chances to pass before next spring, so maybe the problem will go away if a large number of these students pass. If this does not happen, we all wonder what the legislature will do. Will they rescind the functional literacy test legislation? Will they keep the legislation but delay the implementation?

Will they require easier content or passing scores? Or will they maintain the present legislation? Stay tuned to Florida for later developments.

How far should the politicians go in directing or affecting the minimum competency movement? In almost all the states some sort of minimum competency legislation has been passed. Should politicians set the content of the tests? If so, will this result in special interest groups inside education or outside of education campaigning for their own special interests? Will politicians be subjected to pressure to include measures of performance in "old arithmetic" or "modern math," in "civics" or "civil rights," in "Americanism," or "Consumerism"?

Should politicians set the acceptable level of performance on a test? They have, in effect, in Florida. They do not worry, as psychometricians do, about the technical issues involved in setting a cut score. Should elected officials determine extent of test item bias? They do not have the limitations of psychometricians, and do not hesitate to throw items out if they "know" they are biased.

The preceding paragraphs are meant to suggest that some political leaders are getting deeply involved in competency testing, primarily because it is a popular issue with their constituencies. Their involvement may take them beyond mere enabling legislation into more technical areas such as test content and construction, establishing cut scores, and determining item bias. Through expediency, lack of knowledge, or both, political leaders will quickly and boldly make decisions in areas that psychometricians and educators make very slowly and deliberately. The challenge to educators is to educate the political leaders in this area,

90

and to take leadership roles in solving these problems. If we do not take a leadership role, others certainly will and the results may be unsatisfactory to many of us.

## Measurement Issues

We still need to be concerned about the adequacy of the cut score. There is the argument, voiced by Gene Glass (1978), that there is not a defensible method for setting a cut score, and that group processes for doing so are inexcusably unreliable. There are others, who say reliable judgments can be made by groups of experts (Zieky and Livingstone, 1978). The debate on this issue goes on.

There may be problems in setting a cut score, but there are also problems in not setting a cut score. Taxpayers will not be happy to learn that the education profession is not able to judge the adequacy of student performance. Someone needs to set a level of performance, because one will be set. It is better, I believe, that this be done by an informed educational community than by other groups who are not experienced and who may make their decisions with self-serving or other motives.

Validity is an issue of primary concern to any conscientious testing project. How will validity be established for minimum competency tests? We believe the most important validity to be established for an objective referenced minimum competency program is content validity. The content validity will be the degree to which the items measure the skill specified in the objectives. This is accomplished through applying human judgment to determine if there is appropriate consistency between the objectives and the test items. If the item adequately measures the objectives, an important validity issue is solved.

An equally important problem is to validate the skills on which the tests are based. Conscientious educators and informed laymen must be involved in helping to select and specify the skills to be measured by the minimum competency test. Establishing the validity of the skills is equal in importance to the validity of the test items.

Differential performance by groups, when it occurs, is a problem for a minimum competency test program. On the first administration of Florida's Functional Literacy Test, the failure rate was 77% for the black students and 24% for the white students. The concern for bias in the tests under these circumstances is understandable and careful study should be conducted. Human judgment is probably the most important factor in identifying and neutralizing bias in tests. (Hills and Maruhnich, 1977).

It should also be recognized that differential performance is not synonymous with bias. Differential performance may indicate differential level of skill or true functioning. The problem, as always, will be to distinguish between differential performance due to test bias and differential performance due to differences in true functioning level. If a test is picking up bias, it should be discarded. If a test is measuring, fairly, what it is supposed to, it needs to be retained even in the face of criticism. It would not be appropriate to discard a scale simply because it showed a certain group of babies to be dangerously underweight. If the scale is fair and accurate, the solution is feeding the babies, not destroying the scale.

This brings up another characteristic important in competency testing: modesty in interpreting the results. With any testing there are going to be some examples of misclassification. Drastic and irrevocable decisions should not be made based on limited evidence, such as a single test score.

There is a narrow line to be walked by those responsible for a minimum competency testing program. There must be teeth in the testing, or the results will be ignored. But the consequences can not be so irrevocable that misclassifications cannot be reversed. There are costs to both students and society, if the student is judged a non-master when he is a master, or if the student is judged a master when he is a non-master. We need to do all we can to eliminate misclassifications.

In Florida, we effected this by allowing students to retake the Functional Literacy Tests, and by providing opportunity for "override" of the results of the basic skills test at the local level. The local school system is also responsible for remediating and certifying mastery of the skills for students who do not pass the basic skills test the first time. Special provisions are available for students who have special test administration needs, such as large type tests, extended time, or a special setting.

## Educational Issues

There are educational problems that need to be solved. Have we really identified the most important, right competencies? What is the criteria for judging their importance or rightness? Should they be judged on their simplicity or on the need to know the skills they test in the grade after this grade or in the life after school? This is the problem of validity of the objectives mentioned earlier.

Are we going to end up with only trivial objectives, with emphasis placed on teaching only trivial skills? Some basic skills are hard to measure in a multiple choice format. Will we only test, and therefore, only teach, the skills that are easy to measure?

Are we going to test students, identify those who have not met a minimum level of performance, and stop there? It is not humane for us to deny promotion without providing effective ways for the student to learn and demonstrate proficiency in the skills in which he is deficient. This requires remediation efforts and additional testing opportunities for the student.

Will we attempt diagnosis and prescription from such tests? Or will they only signal areas that might need further examination? There is a tendency to overinterpret simple test results.

Can we teach the students who fail, or will they be permanent residents of remedial education programs? We have always been able to teach the bright student. Many of them would learn without the intervention of a public school. Students at the low end of the ability scale have been placed in special classes with special instruction and with different expectations. But students who are not eligible for special education, but are still toward the low end of the ability scale have been hard for us to teach. Do we really know how to teach these students? Are social promotion or retention the only options available for these students? I believe the problem of structuring curriculum to teach these students is the most difficult educational problem we face in minimum competency testing.

The high expectations of the general public for improved school systems may also create problems for education. The CBS special "Is Anyone Out There Learning," broadcast on August 22, 23, and 24 described "minimum competency" and "back to basics" as a grass-roots movement to improve the school systems which have lost the public's affection. The program went on to describe the many problems that existed in school systems today. The problems included racism, student apathy, drugs, inflation, and parental abdication of responsibility for teaching their children. It is clear that minimum competency testing cannot solve all those problems, for they are societal problems. We must continually focus on the narrow role minimum competency testing can play: identifying academic deficiencies in discrete areas of the school program.

I believe there are legal, philosophical, political, measurement, and educational problems to be solved if minimum competency testing is to continue. Florida has faced some of these problems and has survived. The driving motivation behind Florida's minimum competency program is to identify students who have not acquired certain skills and to teach those skills. We believe it is our responsibility to do this, rather than promote students only because they are one year older. We believe that the best minds in education will be able to solve the problems suggested in this paper. We believe minimum competency testing, by identifying student deficiencies so that appropriate action can be taken to resolve and avoid further reoccurrences of those deficiences, will continue and be beneficial to public education.

# REFERENCES

Glass, Gene B. "Minimum Competence and Incompetence in Florida," Kappan, 1978, 59, 602-604.

Hills, John R. and F. J. King. "Construct Validity of the Florida Functional Literacy Test," unpublished paper produced under contract to the Florida Department of Education, 1978.

Hills, John R. and Nick Maruhnich. "Test Item Bias: An Analysis of Methodology and Recommendations," unpublished paper produced under contract to the Florida Department of Education, 1977.

Northcutt, Norvell. Adult Performance Level Study. Austin, Texas: the University of Texas at Austin, Division of Extension, January 1973.

Zieky, Michael J. and Samuel Livingston. Manual for Setting Standards on the Basic Skills Assessment Tests. Princeton, New Jersey: Educational Testing Service, 1978.

# PROBLEMS AND PROSPECTS IN MINIMUM COMPETENCY TESTING*

Walt Haney
Project Director
Huron Institute

Competency testing clearly seems to be a major movement with serious implications for American education. According to the Education Commission of the States (ECS) which has taken on the role of chronicler of the competency testing movement, a majority of the fifty states has already taken some sort of action concerning competency testing. In a dozen states, legislation mandating competency testing has already been passed. In another twenty states, state boards of education have adopted regulations requiring some form of competency testing (Pipho, 1977, 1978a, 1978b). The ECS reports that "Minimal competency testing for high school graduation and grade-to-grade promotion continues to be one of the most explosive issues on the educational scene today" (Pipho, 1977, p.i). However, this enthusiasm for competency testing poses a curious contradiction; it comes at a time when more and more questions and criticisms are being raised about tests and test use (See Houts, 1977). In light of this contradiction, it is not surprising that predictions regarding the impact of minimum competency testing on American education have varied markedly. In this paper, I review several of the more extreme predictions for the future of this movement and point attention to five issues which will likely prove crucial in determining whether the promises or the problems of competency testing predominate.

Prognoses for the future of competency testing range widely. Robert Ebel, arguing the case for minimum competency testing, predicts:

---

* This paper has been adapted from a longer article by Walt Haney and George Madaus to appear in the Harvard Educational Review in fall 1978. Work on the paper was supported by the National Consortium on Testing with funding from the Carnegie Corporation of New York. Views expressed in the paper do not, of course, necessarily represent those of anyone other than the author.

> Tests of minimum competency are likely to actually increase achievement, particularly for pupils who seem less interested and successful in learning. Such tests can motivate efforts to learn and efforts to teach by providing specific goals and incentives (Ebel, 1978).

From a very different point of view, Gary Hart, California Assembleman and former teacher, suggests that minimum competency testing:

> can accomplish three goals: create a public dialogue regarding high school graduation standards, restore meaning to the high school diploma, and insure that students who have difficulty with basic skills receive appropriate instruction (Hart, 1978).

Minimum competency testing, of course, has its critics as well as its enthusiasts. Arthur Wise calls such testing another case of hyper-rationalization--that is, a persistent effort to rationalize despite ignorance of the relationship between means and ends. He observes that "reason and evidence provide little, if any justification for the belief that minimum competency testing will help poor students to learn or poor teachers to teach" (Wise, 1978). Elsewhere the competency testing movement has been depicted as a fraud:

> It's like the bait-and-switch advertising tactic, in which at the point of sale an inferior item or less attractive product is substituted for the more desirable item. The attractive product is an education that develops competencies needed for life....The switched product in the bait-and-switch swindle is a score on a standardized paper-and-pencil competency examination. It is a poor substitute for the real thing: performance at life's tasks (Nathan and Jennings, 1978).

Reactions have varied not just to minimum compentency tests in general, but also with respect to specific cases of such testing. For example, in a meeting in May, 1978, Florida officials predicted that their state's minimum competency testing program was a harbinger: it would be adopted in other states sooner than educators expect (Report on Education Research, May 17, 1978). In the same month, however, Gene Glass predicted in the Phi Delta Kappan that Florida's minimum competency graduation assessment "will be suspended before the end of the year" (p. 603).

Why have reactions to the minimum competency testing movement been so varied? After all, the idea seems a natural extension of other themes in social policy--legislated minimum wages, for example, and mandates for minimally adequate health care. Why, then, should we encounter such debate about proposals to guarantee minimum levels of learning?

It is, of course, impossible to be sure why a movement like minimum competency testing finds public favor. The rapid rise of the minimum competency testing movement seems due largely to the fact that many diverse interests have converged on the idea. Likewise, a whole range of fears seems to have focused on the issue of competency testing. Such fears and hopes for competency testing often seem to have as much to do with economic and political affairs as with strictly educational concerns. Since I cannot here delve into the ranges of economic and political perspectives bearing on competency testing, let me focus on five issues which are more clearly educational matters and which I think hold special relevance as to whether or not the promises or the problems of competency testing predominate.

First is the relationship between what is tested and what is taught. In the rush to test for competencies, remarkably little attention has been given to how the competencies or skills tested relate to present school curricula. The match between what is tested and what is taught needs considerably more attention. Previous efforts along these lines have shown that the correlations between what is tested in schools and what is taught may be lower than is often assumed (Jenkins and Pany, 1976).

A provocative example of this problem was recently made available by Case Western Reserve School of Medicine. For more than twenty years, this school used a standardized medical examination to determine whether or not students should be promoted into the third year of medical training. Faculty at this school became concerned, however, when it was shown that many students could pass the examination without ever having taken the second year's course of study. When the examination and the medical curriculum were compared, it turned out that much of what was asked on the examination was not taught in the second-year curriculum and that significant areas that were taught were not tested. As a result, the Case Western Reserve medical faculty abandoned the standardized examination and developed another that corresponded better with the curriculum (Wile, 1978).

The problems confronting American public education are, of course, quite different from those faced by the Case Western Reserve medical school. A recent CBS news special, for example, posed the question, "Is Anyone Out There Learning?" (CBS News, 1978). The public seems to have answered "no," or at least, "not enough of what should be learned." The trouble with competency testing as a response to such reaction is that in the rush to test, very little attention is being given to the issues of what should be learned that is not now being learned and what should be taught that is not now being taught.

It just may be that teaching and learning are weakest in the areas where traditional standardized testing is weakest. The National Assessment of Educational Progress suggests, for example, that between 1971 and 1975, the reading skills of 13- and 17-year olds remained fairly stable. On literal comprehension, the performance of 17-year olds remained the same; for items dealing with inferential comprehension, there were slight declines in performance of 17-year olds (ECS, 1976). A later national assessment showed, however, that between 1969-70 and 1973-74, there was an overall decline in the coherence and quality of essays written by 13- and 17-year olds (ECS, 1977-78). A recent nationwide study corroborates these findings. A variety of groups surveyed--including students, teachers, administrators, and legislators--reported that students have more difficulties with writing than they have with reading or mathematics. Respondents also rated general student proficiencies as being weaker in speaking and listening than they were in reading or mathematics (Bossone, 1978).

The results raise real questions about the extent to which schooling should focus on these various skill areas. But whatever the case, such questions should not be left to be decided merely on the basis of what can be conveniently tested. If the educational goals embodied in competency testing schemes are not adequately covered in current curricula, the discrepancy needs resolution. Testing students on what they have not been taught is unfair and denying them a high school diploma on that basis may well prove illegal. If the match between what is tested and what is taught is not good, testing may also affect the curriculum in unintended ways. Experience with programs using tests to certify successful completion of a given level of education shows that examinations quickly influence what is and is not taught, how it is taught, and how pupils study (Madaus and McNamara, 1970). If students are denied high school diplomas on the basis of minimum competency examinations, these tests could quickly become a powerful influence on what is taught and learned in American schools.

A second key issue in competency testing is the idea of minimums. Much recent debate has centered on the appropriate methods by which minimum levels of competence might be determind (Zieky and Livingston, 1977; Glass, 1978; French, 1978). Thomas Green (1977) has argued that "goals that we express in policy are always minimal standards because in the social practice of a system--as opposed to the theory of its management--they operate as measures of failure rather than as standards of success" (p. 3). Recent experience does seem to support Green's assertion. Competency tests, more frequently than not, seem to have become minimum competency tests. Nevertheless, the necessity of emphasizing minimums is not altogether clear. Indeed, in March, 1978, Ellis Page, president-elect of the American Educational Research Association, proposed a scheme for completely skirting the minimum issue. He argued that a system of scaled certification, "the reporting of a brief profile of scores from well designed senior tests," would be a way of escaping the problems of setting and defending minimums while at the same time adding meaning to the high school diploma (Page, 1978).

Page's logic seems undeniable. Such a scheme would do away with debates about appropriate minimums. Yet the Page proposal, unlike minimum competency testing, has not caught on. The reason for the continuing emphasis of minimum competencies, despite heated debate on how they reasonably can be established, may be due to a variety of factors. Perhaps, as suggested by Green, social welfare guarantees must in a highly competitive society focus on social minimums if they are to avoid challenging fundamental assumptions about how social goods are allocated.

One possible educational explanation of the minimalist orientation of competency testing has to do with the idea of remediation--a third issue to which I would like to draw attention. Specifically, it might be argued that we need to establish minimum levels of performance so that testing can serve as a trigger to remediation efforts when results indicate that minimums have not yet been attained. The need for remediation efforts for those not meeting the minimums seems implicitly recognized in the way competency testing schemes generally are organized. Most minimum competency testing programs quite sensibly have not waited until the twelfth grade to administer tests to students. Students are usually given several chances to pass them, beginning early in their high school careers. Relatively little attention has been given, however, to exactly how remedial help can best be given to students who do not pass competency tests the first time around. The situation in Florida seems prototypical of this problem. Only after nearly forty percent of eleventh graders taking the mathematics proficiency test failed, did the state "provide a compensatory education program that distributed $10 million to local districts (on the basis of state assessment scores) to provide remedial instruction" (Fisher, 1978).

The availability of well thought out remediation efforts on a broad scale--and our experience with compensatory education at the elementary level should give us pause here--will be crucial in determining whether minimum competency testing programs really help to improve learning. If not, they may simply supply a new means for tracking weak students into second-class educational programs and stigmatizing them as inferior citizens.

This leads us to a fourth issue, the potential of competency testing for unfair discrimination. Disadvantaged students, including minorities, handicapped children, and students who do not speak English as their first language, typically have not had educational opportunities equivalent to their middle-class peers. Therefore, the potential is great that minimum competency tests will have an adverse impact on these groups. Already this issue has been raised in several southern states. In North Carolina, a boycott of the statewide competency test has been advocated by minority leaders (Chatham County Herald, July 19, 1978). In Florida, the NAACP has announced its intention to file suit to enjoin further administration of Florida's proficiency examination on the grounds that it is culturally biased (Education Daily, March 29, 1978). Regardless of whether the tests themselves are biased or whether minorities have received inferior instruction, the potential for discrimination seems great. As a recent Southern Regional Council report suggested, tracking based testing may become the new segregation (Mills and Ryan, 1976).

This brings us back again to the nature of competency tests, and with respect to the tests, I would like to raise a fifth issue, namely that of test security and freedom of information. Airasian, Madaus, and Pedulla (1978) observed that one problem little noted in discussions of minimum competency testing is the issue of test security. Some minimum competency tests are modeled after the SAT, the Law School Admissions Test, and other completely secure tests. In such cases, the same tests, or large portions of them, are used year after year and are never made public. For tests such as the SAT and the LSAT, this approach makes some sense; individuals voluntarily take the test and pay the examining body a fee for the services they provide in making up and administering them.

The completely secure model of testing does not, however, seem appropriate for minimum competency testing. First of all, students will not take such tests voluntarily, instead they will be required to do so. Second, state and locally sponsored tests are being developed with public monies, and freedom of information laws may require that they be made public after administration. If they are, test security will vanish, and new tests will have to be developed for each administration. External testing programs tied to high school certification have in the past followed this procedure rather than the totally secure approach of the SAT. Following this tack, however, will increase the cost of a minimum competency testing program because test development will not be a one shot affair (Airasian, Madaus, and Pedulla, 1978).

The issue of freedom of information concerning standardized testing seems likely to grow far beyond the bounds of minimum competency testing.

The 1977 version of the American Psychological Association's Ethical Standards of Psychologists has already decreed:

> Persons examined have the right to know the results, the interpretations made, and, where appropriate, the original data on which final judgments were based. Test users avoid imparting unnecessary information which would compromise test security, but they provide requested information that explains the basis for decisions that may adversely affect that person or that person's dependents (APA, 1977, p. 6).

Where do test takers' rights to information leave off and needs for test security begin? The proliferation of tests developed with public monies and required for high school graduation is likely to spur more attention to the rights of individual test takers, especially when tests are intended for purposes of instruction rather than for personnel selection.

In sum, the five issues I have mentioned suggest to me that the problems of implementing minimum competency testing programs appear at present to be far larger than the promises of such programs. The major problems concern the relationship between what these tests measure and what is actually taught in schools. This relationship, or lack of it, deserves considerably more attention than it has so far received. If the relationship is weak, the tests are unfair; if it is strong, it may be because what is tested affects what is taught. Competency testing schemes may have real--though unintended--consequences, both for students and for the substance of what gets taught in school. The most sensible antidote to such unintended consequences is to devote more attention to what should be taught in the schools and how to teach it more effectively, and less on merely that which conveniently can be measured. For the minimum of what currently can be measured, especially using paper-and-pencil competency tests, should not define the minimum of what is taught or learned.

## REFERENCES

American Psychological Association. Ethical Standards of Psychologists (1977 Revision). Washington, D.C.: APA, 1977.

Bossone, Richard M. Minimum Competencies: A National Survey. New York: Center for Advanced Study in Education, 1978.

"Competency Test Boycott Proposed." Chatham County Herald, July 19, 1978.

Education Commission of the States. Reading in America: A Perspective on Two Assessments. Denver, Colorado: 1976.

Education Commission of the States. "Writing Skills--Mechanics Expression, Persuasion and Revision." Denver, Colorado: 1977-78 (Summary).

Education Commission of the States. "Policy Issues in Minimum Competency Standards and Issues." Denver, Colorado: April 19, 1978 (Draft).

Ebel, Robert. "The Case of Minimum Competency Testing." Phi Delta Kappan, 59 (1978), 546-548.

Fisher, T. H. "Florida's Approach to Competency Testing." Phi Delta Kappan, 59 (1978), 599-602.

"Florida's Minimum Competency Testing: Is it Coming or Going?" Report on Education Research, May 17, 1978.

"Florida NAACP Will Sue State Education Officials Over Literacy Tests." Education Daily, March 29, 1978.

Fremer, J. "In Response to Glass." Phi Delta Kappan, 59 (1978), 605-606.

Glass, Gene. "Minimum Competency and Incompetency in Florida." Phi Delta Kappan, 59 (1978), 602-605.

Green, Thomas. Minimal Educational Standards: A Systemic Perspective. Background paper prepared for ECS/NIE/Carnegie regional conferences on minimum competency testing, 1977.

Hart, Gary. "The California Pupil Proficiency Law as Viewed by its Author." Phi Delta Kappan, 59 (1978), 592-595.

Houts, P. (ed.). The Myth of Measurability. New York: Hart, 1977.

"Is Anyone Out There Learning? A CBS News Report Card on American Public Education." Transcript of a broadcast over the CBS television network. New York: CBS News, August 22-24, 1978, Parts I-III.

Jenkins, J. and D. Pany. "Curriculum Biases in Reading Achievement Tests." Urbana, Illinois: Center for the Study of Reading, University of Illinois at Urbana-Champaign, November 1976.

Madaus, G. and J. McNamara. Public Examinations: A Study of the Irish Leaving Certificate. Dublin, Ireland: Educational Research Centres, 1970.

Mills, R. and M. Bryan. Testing...Grouping: The New Segregation in Southern Schools? Atlanta, Georgia: Southern Regional Council, 1976.

Nathan, J. and W. Jennings. "Educational Bait-and-Switch." Phi Delta Kappan, 59 (1978), 621-624.

Page, E. B. "Escaping the Minimum Competency Dilemma: Scaled Certification." CAPT, Clearinghouse for Applied Performance Testing Newsletter, 3, No. 5 (1978), pp. 6-12.

Pipho, Chris. Update VIII: Minimal Competency Testing. Denver, Colorado: Education Commission of the States, November 15, 1977.

Pipho, Chris. "Minimum Competency Testing in 1978: A Look at State Standards." Phi Delta Kappan, 59 (1978), 585-588. (1978a)

Pipho, Chris. "State Activity Minimal Competency Testing." Denver, Colorado: Education Commission of the States, May 18, 1978. (1978b)

Wile, M. "External Examinations for Internal Evaluation: The National Board Part I as a Case." Journal of Medical Education, 53 (1978), 92-97.

Wise, Arthur. "A Critique of Minimum Competency Testing." A paper prepared for the Educational Adequacy Program Basic Skills Group. NIE, September 1977.

Zieky, M. and Livingston. Manual for Setting Standards on the Basic Skills Assessment Tests. Princeton, New Jersey: Education Testing Service, 1977.

# PART IV - STATE AND FEDERAL ROLES IN TESTING: RISE OR DECLINE

# THE STATE ROLE IN TESTING:
# AS VIEWED BY A CALIFORNIA EVALUATOR

Alexander I. Law
Chief
Program Evaluation and Research
California State Department of Education

Education is being asked to prove its worth, and this proof is invariably tied to some type of testing, usually normative in character. That is, people believe that test scores are a reliable measure of educational progress, a belief evident in the furor accompanying both the decline in SAT and national reading scores and the failure of many twelfth graders to pass minimum competency tests. Large-scale evaluations of various federal and state educational programs, such as Bilingual, Title I, and Follow-Through, have only added to the furor. Reports issued by firms funded by the Federal government to evaluate such programs have been highly critical. Needless to say, such reports are grist for the political mills, both federal and state. And criticism usually precipitates action. The action stimulated by such negative reports has not been to terminate these programs, but to exhort federal and state evaluative agencies to shape up and get more test scores.

Certain congressmen, Senator Pell, in particular, are now pushing for national tests and standards of competency. The National Academy of Education advises against such a move. Although these actions and reactions do not necessarily portend more testing, they certainly focus on testing as a primary means of evaluating educational programs. The national concentration on minimum competency testing on both the high school and grammar school levels is a case in point.

At the present time, 36 states have some type of legislation regarding competency testing. Of these 36 states, only 9 delegate the setting of standards to local agencies. In contrast, 25 require that their state departments, boards, or commissions of education establish statewide standards of competency. One state is undecided as to whether standards should be set by a central agency or by local agencies, and one other state requires its Department of Education to set graduation standards and permits its local agencies to set standards for testing in elementary grades. That only 9 states permit the setting of standards to local educational agencies seems to me a headlong plunge into a centralist position and augers for a repetition of the problems encountered in Florida, one of the 25 states which have established central rather than local standards of competency. Although a discussion of the Florida problems is beyond the scope of this paper, they have been indirectly touched upon in Mr. Loewe's presentation and have been extensively written about by Gene Glass.

California is one of those states which have instituted competency testing. We believe, however, that standards should be locally determined. The California Legislature has mandated that the State Department of Education provide assistance to local educational agencies, via assessment devices, information on staff development, and various educational materials. But the final judge of competency standards and of the assessment process will be local governing boards. While the testing is to be accomplished by local agencies, the impetus for such testing was initiated by the state legislature. The stimulus for such testing was, of course, the same as that in other states: the perceived inadequacy of many students in the basic skills. This concern raises some interesting questions as to the definition of basic skills and how these skills should be assessed. A quote from a letter I received from NCTE this week exemplifies the issue:

> "The XYZ Unified School District's list of minimum competencies was drafted by a committee representing students, parents, graduates and staff members, and the cooperativeness of the effort is commendable. Less commendable is the fact that the writing competencies are confined to six skill areas--spelling, punctuation and capitalization, standard forms (the filling out of, I presume), sentences, paragraphs and handwriting. Would that the competencies had included at least minimal demonstrations of a few thinking processes, e.g. categorizing or comparing, or performance objectives to indicate some basic concepts or attitudes, such as an awareness of the cultural diversity of our society. To those who may say, "But this is a writing test," I would reply, "Can writing be divorced from thinking?" Apparently the makers of this test think that it can be. The reading taxonomy does lend itself to the possibility of testing for thinking and attitudes, but the sample test on reading is confined to functional literacy. Furthermore, I should like a test on minimum competencies to include performance objectives for speaking and listening as well as reading, writing and math."

In other words, there has to be a consensus as to what should be tested before we can have meaningful competency assessment.

The problem of determining the basic skills and deciding how they should be tested leads us to three interrelated topics: (1) the use of testing to improve instruction; (2) changes in evaluation techniques; and (3) trends toward fiscal conservatism as it relates to school resources.

With respect to the first topic--the use of testing to improve instruction, let me note that we have spoken about the competency, proficiency, and accountability thrusts, all of which call for an increase in testing or a change in the mechanical aspects of score reporting. But none of these thrusts focus directly on what should be the major issue in testing: that test scores lead to more appropriate instructional efforts and higher student achievement. And a change in test reporting will do precious little to either improve competency assessment or to generate student achievement. Our primary concern, then, should be to determine the sensible relationship of information gained from test scores to a sound curricular approach. Certainly, the proliferating aggregation of federal and state test scores offers little, if anything, to participating schools in terms of improving or reshaping curriculums. The schools may, of course, get some information from the norm reference test, but unless school evaluators are remarkably skilled, this is not likely. Unfortunately, schools and school districts can easily dissipate their energies in the test data game, rather than conserving their energies for more important work: the determination of how testing can improve curriculums and effect higher student achievement.

With respect to the second topic--changes in evaluation techniques, we should note that many of the states require remediation as part and parcel of competency testing. If evaluation techniques are sound, specific student disabilities can be ascertained and corrected. But when we view testing in this larger context of evaluation, it seems that there has been an historic over-reliance on scores and other quantitative data. Early evaluation efforts were modeled on classical experimental designs, designs which appear to be not only impractical but impossible to implement in social or educational programs. Let me emphasize that

I am not questioning the utility of psychometric information if used properly. What I am questioning is the increased reliance at both state and federal levels on test scores alone. Here in California, we have found that case studies and ethnographic information provide greater insight into the value of government funded programs than simple tests of statistical significance. Although the movement from the quantitative to the qualitative seems to be gaining strength in other parts of the country as well, we do not want to become polarized into an either/or situation. For more successful evaluative techniques, we need to use all available sources of information, both quantitative and qualitative.

My third topic--fiscal conservatism (my term for what is variously called the taxpayer's revolt or, locally, Proposition 13), is also related to the whole question of testing. It is my experience that when state resources are used to fund local educational programs, as must be the case when local financial support is withdrawn, control follows funding. With legislative control comes legislative inquiry into the relative merits of competing educational programs. With limited resources, the legislative budget committee must ask which programs should be funded and which should be curtailed. Legislators insatiably seek information, information which they believe they understand, and test scores have a presumed validity and universal appeal to legislative bodies. It seems probable then that, as local resources are replaced by state resources, testing will increase, not as an instructional tool, but as an accountability device.

Earlier, I suggested that trends in testing are tied to politics. Let me reinforce this suggestion by noting that testing programs, their

115

uses, interpretation, and acceptance, are a political act, political in the sense that an act creates change. Testing, in its broadest sense--that meant by Jim Block in his definition of tests as that which makes the unobservable, observable--will increase. We can only hope that this increase will effect better evaluations and, concomitantly, better instruction for those being tested.

# STATE AND FEDERAL ROLES IN TESTING:
# AS VIEWED BY THE STATE SUPERINTENDENT OF NORTH CAROLINA

A. Craig Phillips
Superintendent of Instruction
North Carolina Department of Education

To begin, let me say it is not my purpose to debate the merits of achievement and competency tests. A comprehensive testing program in most states is now a reality. According to the latest figures I have, thirty-six states have now mandated statewide achievement testing programs, and most of the other states appear to be moving in this direction.

Unlike many programs that are introduced in public education, the movement toward statewide achievement and competency testing is being initiated and supported by the lay public rather than by professionals from the education community. Furthermore, there is ample evidence that the interest and demand on the part of the public will not dwindle and pass away. To the contrary, there is reason to believe that this insistence on educational accountability will increase and become permanent in nature.

We are told by no less an authority than Alexander Pope that the best program is that program which is best administered. While such a generalization falls short of absolute truth, it is not entirely without substance. A testing program, wisely administered, can be both a sensitive aid to the teacher in providing personalized attention to a student's academic needs, and an aid to the local administrator in evaluating the success of his instructional programs. It should also serve as an important source of information for the legislator in allocating educational funds and as a means of accountability to the general public. On the other hand, a poorly administered testing program can become a harsh, repressive instrument, freezing children in their places, discouraging efforts by teachers, generating defensive and evasive behaviors from administrators, disillusioning legislators, and creating a hostile public. Since we no longer have the option of weighing the dangers and deciding whether or not to initiate testing programs--testing programs are here to stay--we must be doubly sure that the benefits of testing are obtained and its dangers avoided.

I for one--and I am sure that there are many other educators who concur--believe that a publicly initiated movement in testing is encouraging. When, in the past, have we had such an interest and involvement in public elementary and secondary education? When, in the past, have we had such a demand for quality--in terms of student performance--from the consumer of the educational process?

Therefore, as a school officer in a state that has just begun comprehensive mandated statewide achievement testing in grades one, two, three, six, and nine and a competency testing program, as one requirement for graduation, in grade eleven, I am committed to use these testing programs as tools to improve instruction and the management of education and, ultimately, to enhance the well being of the young people who attend the public elementary and secondary schools in North Carolina.

Although the United States Office of Education (USOE) has a definite role in helping North Carolina and other states with their testing obligation, the major responsibility should rest with state education agencies. Therefore, the role of the federal government in educational testing should be limited to support activities. First, and foremost, there should not be a national testing program. It is difficult enough for a statewide testing program to provide for the unique needs of local communities; it would be impossible for a national program to accomplish this task. USOE. through the National Institute of Education (NIE) or other subdivisions, could, however, make a significant contribution by conducting and supporting such activities as:

Promoting and funding research and development activities in testing;

Recommending, not mandating, standards for testing programs;

Providing technical assistance to states;

Collecting, screening, and disseminating information about testing and evaluation needs to the United States Congress.

The role of the states in testing, however, must be more specific and direct. This is true for two reasons. First, the state agency is closer to the grass roots and can best serve the unique needs of the schools under its jurisdiction. Secondly, state constitutions and legislated testing programs mandate that the state agency accept leadership and coordination responsibilities relative to implementing state testing programs. For example, in North Carolina the testing legislation is explicit in terms of the roles and responsibilities of the State Board of Education, the Chief State School Officer, the State Education Agency, the local school boards, and the local education agencies. In our state, major leadership, coordination, and quality control activities have been assigned to the state agency. Even if these functions were not mandated to the state agency, in my opinion, it has a professional, as well as a moral, obligation to assume such responsibilities.

In the limited time I have been allotted, I would like to review some of the state's responsibilities in implementing achievement and competency testing programs. These responsibilities can be logically grouped under four topic areas.

1. The selection of proper measuring instruments--basic to the success of the entire program;
2. Accountability--helping the public understand how well the educational system is performing;
3. Using information for management and instructional decisions and actions; and
4. Providing the resource for implementing the testing program and correcting deficiencies.

## Test Selection

First, the state must provide leadership in the selection of testing instruments that, to the greatest degree possible, measure the objectives of the instructional programs being implemented in local school districts. Proper test selection assures valid accountability as well as pertinent information for learning diagnosis and prescription. Working with test publishers, state agency personnel must involve teachers and other instructional leaders in identifying instructional objectives and then selecting test items and tests to measure what is being taught. Every caution must be taken to insure that the tests measure what should be and is, in fact, taught in the classroom. We must not allow test content to dictate instructional objectives and curriculum content. We must demand that instructional objectives and curriculum content dictate test content. Testing should follow naturally in the learning process. Testing determines the existence, the quality, and the genuineness of learning. (In North Carolina, two testing commissions, working with state agency personnel and local teachers and instructional leaders, have and will continue to work diligently to insure that all tests--achievement and competency--have content validity and reflect the instructional objectives of the state.)

The state must also take a dominant leadership role in insuring that test content is free of ethnic, sex, and cultural bias. If a single test item discriminates unfairly against any subgroup in a student population, that test item should not be included in a test. (In North Carolina, state agency personnel, working with two testing commissions and representatives from various populations in the state have devoted much time to eliminate test items that place any particular group of students at an unfair disadvantage as they participate in the competency and achievement testing programs.)

## Accountability Function

The state education agency has a definite responsibility for providing leadership in accounting for the results obtained from educational efforts implemented throughout the state. Time does not permit a complete discussion of this important function, but I should like to make general reference to some of the most important responsibilities of the state relative to accountability.

1. We must insist that all available test data at the state and local level be public information, but, if, in our opinion data are misused or misinterpreted, we have an obligation to bring these discrepancies to the attention of the offenders and the public.

2. The state must issue state reports and provide technical assistance to local school systems issuing local reports if school boards and the general public are to have valid and understandable interpretations of test results.

3. The state must insist that when comparisons of test results between North Carolina and the nation or between school systems and schools within a unit are made that these comparisons be done on a fair and valid basis. This means that the state must insist that norms from the publishers are based on representative samples and that the student composition of the sampling be interpreted in a complete and valid manner. Likewise, the state must develop state norms that will provide the local school systems and local schools with the opportunity to compare themselves with similar school systems and schools in terms of characteristics that are related to student achievement. This is not an effort to provide a cover up for a state or for a school system or a school within a state which might have low test scores in comparison with some norm group. The purpose is to insure that any organizational unit, whether it has a superior or inferior test record, be compared with organizational units with similar characteristics. You know, as well as I, that it is entirely possible for the lowest ranking school system in a state in terms of test results to have a better educational program than higher ranking school systems if the

potential of the school systems is taken into consideration. The administrators and teachers in an organizational unit should not be held accountable for performance in terms of absolute scores; they should be judged rather in terms of how well they do with the resource and types of students they have. (In North Carolina we are working on developing similar school scores and emphasizing that test scores should be interpreted in light of a number of factors that are related to student achievement.)

4. The state needs to help local school systems and the public to think of accountability in terms of progress over time rather than solely the status of achievement at a particular time. For example, it is far more important and revealing for a school system to report that the achievement of a group of students has improved significantly over the last five years than to indicate that the average student in the school system was two months above grade level in 1978.

Using Test Information

In the past, too many of us in education have been guilty of administering tests, writing a summary report, and then allowing the report to gather dust. Test results, at best, have made little impact on management and/or instructional practices. State agencies have the responsibility of seeing that test data is used to the fullest extent, in establishing policy, priorities, long-range planning, in evaluating programs, and improving instructional methods and student performance.

Briefly, some examples of how test data can be used are as follows:

The test report that is presented to the state board of education or to local school boards might conclude with some emphasis on policy or budgetary considerations. For example, in our first annual testing results, male students, especially those in the ninth grade, were far below the achievement level of their female counterparts. The adoption of special programs and/or the allocation of additional funds to attack this problem represent responsible use of test results. Incidentally, the identification of the problem might also suggest a need for additional analysis and research to determine cause.

Test results, along with other information, might also be used by the state agency to establish priorities for such things as requesting resources and developing staff development programs. In the annual testing program, we are obtaining detailed information on the percentage of students in the school, school system, and at the regional level who master and do not master instructional objectives emphasized at various grade levels in North Carolina. It is the responsibility of the state agency to identify instructional deficiencies and to attack these deficiencies through staff development activities.

The state agency has the responsibility of using test data and other information in helping local school districts develop comprehensive plans. State accreditation of local school systems in North Carolina requires that the local school system develop and implement a comprehensive plan. This plan--which requires a needs assessment, the development of priorities and specific objectives, the formulation of appropriate strategies, and evaluations of efforts--cannot be done without the proper use of test information. One of the biggest challenges facing state agency personnel is to become more competent in helping local school personnel use test data in developing plans to meet accreditation standards.

The state agency has a legal and professional obligation to evaluate the effectiveness of all major new programs that are funded with state and federal funds. With a few exceptions, the success or failure of programs will be measured in terms of student performance on standardized tests. Information collected through our annual testing program, coupled with other information, provides us with an information base to meet this obligation. Our task is to become more proficient in evaluating programs at the state level and to provide technical support at the local levels in order that better evaluation can be done by our colleagues.

Finally, the state agency should provide leadership in helping classroom teachers develop instructional management systems that will use diagnostic information for improving student performance. In North Carolina, we are encouraged with the results that teachers of early childhood students have obtained through the proper use of diagnostic test information. We have an obligation to improve and expand on this vital use of test data. Helping hundreds of teachers develop appropriate diagnostic instructional procedures is a mammoth task that requires the cooperation of both local and state personnel as well as the personnel in the teacher training institutions in the state. One of our major aims will be to work with higher education to insure that prospective teachers have the knowledge and skills to use test information for diagnostic and prescriptive teaching prior to the time that they accept teaching positions.

When a state initiates competency testing programs which influence whether a student does or does not graduate from high school, it has special obligations. In addition to selecting valid tests and cutoff scores to insure that students are treated fairly, the state has an obligation to administer remedial programs to correct student deficiencies. These programs must provide for those who fail the tests at the secondary level, but, more importantly, preventive programs must be implemented in the early grades if we are to insure that most of our graduates are competent to function as adult citizens. We are now planning to conduct longitudinal studies at grade eleven, and to use annual testing programs in grades one, two, three, six, and nine to identify those students who will experience difficulty with the high school competency test. Our biggest job and responsibility will then be to provide effective remediation programs.

## Providing Resources for Testing and Instruction

None of the things which I have previously mentioned can be implemented unless adequate resources are available. State agencies, working through the state boards of education, have the responsibility for requesting and lobbying for funds that will support a comprehensive and efficient testing program. Yet, if we stop there, the data collected through the testing programs would be useless. We must provide resources for using test data wisely to improve management and instructional programs. In the final analysis, the value of the testing program, or any other program in public education, must be judged in terms of the positive effect that it has on students. It is toward this target that all educators--at the federal, state, and local levels--should aim as they become more involved in implementing testing programs.

In the past, I must admit that North Carolina and, I assume, other states have not derived full benefits from their testing programs. This happened in part because of apathy and in part because of a lack of technical and professional knowledge and skills. In recent years, however, conditions relative to testing have changed significantly. The computer has given us unlimited capacity to summarize and analyze data; the general public has more interest in accountability and quality education; we have more personnel with training in testing and measurement and computer science; test publishers and educators are more skilled in developing and selecting tests which measure educational objectives; and teachers are becoming more skillful in using test information to obtain maximum academic performance on the part of students. In short, the climate is right for us to use testing as a legitimate and forceful tool to improve the management of education as well as the instructional program. It is the responsibility of the federal government and state education agencies, within well defined roles, to see that this tool is used effectively.

# STATE AND FEDERAL ROLES IN TESTING: AS VIEWED BY USOE OFFICERS

John W. Evans
Assistant Commissioner for Planning, Budgeting, and Evaluation
United States Office of Education

David M. Shoemaker
Project Officer
United States Office of Education

We have been asked to address this conference on the problems of testing and, in particular, to address the question: Is there going to be a rise or a decline in federal and state testing activities? We can answer very simply: Yes, there is going to be increased testing activity, although of quite different kinds, at both federal and state levels.

Thus, we would agree with the conclusion that emerged from last year's conference, that there is not going to be a major decline in testing, nor do we believe there should be. In our judgment, testing and test-related activities at both federal and state levels will increase in the coming years, simply because carefully-constructed and properly-used tests provide us with highly valuable knowledge about how children should be educated and how the educational enterprise should be operated. The irresistible conclusion that the values of tests far outweigh their deficiencies will overcome the radical proposals for reducing or eliminating tests.

This general conclusion does not mean that we are unmindful of the serious shortcomings of many tests. Therefore, use of tests, which we believe will increase in the coming years, must be accompanied not only by an awareness of their limitations, but also by a shared commitment by all who develop and use tests to guard against their abuses and to make continual improvements in them. One of the principal conclusions enunciated by HEW Secretary Joseph Califano in a major address on testing before the College Entrance Examination Board last October was that "we need to do more testing and we need to do better testing."

Even though we believe that testing and test-related activities at both federal and state levels are going to increase in the coming years, the two levels of government are going to play quite different roles. And since the different roles of these two levels of government is the principal topic we are addressing here today, let us give a quick overview of what federal and state governments will be doing in the coming years and take note of how their roles in the larger arena of testing are quite different and yet, we believe, complementary.

## State Testing Activities

We see four principal areas in which the current level of testing activity by the states can be expected either to continue at its present level or, more probably, increase.

1. Statewide Assessment. Most states now have fairly extensive testing programs designed to inform educational administrators and the public on the general level of knowledge and achievement being attained by public school students. In the past, these programs usually employed standardized tests which allow comparison of the achievement levels of local students to national norms, but increasingly tests are being tailored

to state and local emphases. It does not seem likely that this type of testing will decrease. Despite the strictures, some of them legitimate, that have been raised and will continue to be raised about these kinds of tests, it is likely that the public as well as educational administrators will want more, not less, of the information these kinds of tests produce.

2. Minimum Competency Testing. Thirty-six states now have minimum competency testing programs of one kind or another, and most of the remaining states will soon install them. Though many of these are now aimed only at determining whether high school graduates have achieved some minimally-acceptable level of competence, the use of competency tests is being expanded to validate promotion from one grade to another and to provide for early exit from high school for those who can demonstrate the required competence in the major content areas. Thus, the increase in the number of states adopting minimum competency tests for multiple purposes indicates a general increase in the development and use of competency-type tests throughout the country. Moreover, we can expect that the pressures for improved educational performance and accountability, growing out of the Proposition 13-type movements now sweeping the country, will provide further initiatives for expanding both competency and assessment tests at both federal and state levels of government.

3. Diagnosis. In the past decade, education has moved increasingly toward the philosophy of individualized treatment. The basic paradigm in individualized instruction is to find out what a child's needs and accomplishments are so that an appropriate educational program can be tailored for him or her. Subsequently, repeated diagnostic measurement is made of the child so that the educational program can be changed or phased out as the need dictates. This trend alone suggests that we can expect an increase throughout our educational institutions in the use of individual diagnostic tests. In addition, the new Education for all Handicapped Children law (P.L. 94-142) will result in a major increase in the use of diagnostic tests for handicapped children as schools struggle to comply with the formidable requirements of this law.

4. Program Evaluation. The substantial increase in rigorous, quantitative evaluations of educational programs, going on at the federal level since 1971, is now expanding to state and local governments. Legislators and educational administrators at the state level, as well as the general public, are demanding more objective evaluations of the various state-funded educational programs. Many of these evaluations will require the use of tests; therefore, we can expect the general increase in program evaluation activities to contribute to the increased use of tests.

As we list and analyze these various factors and trends which we believe will result in the increased use of tests, it is important to step aside and observe that testing is not the name of the game. Test scores will not by themselves bring about educational improvement--which is the name of the game. Thus, while test results can provide invaluable knowledge to educators and decision-makers on what needs to be done, the increased use of tests must exist in a context in which educational objectives are more sharply defined and appropriate educational programs are used to deal with the needs and problems that are revealed by the tests.

### Federal Test-Related Activities

Although we believe there will be an increase in federal test-related activities, it is well to begin by noting what the federal government will not be doing. The federal government will not be imposing, recommending, or even developing national tests. This is strongly-felt current administration policy, and we do not think it is likely to change in the foreseeable future for several reasons:

1. The federal government does not wish to upset the established predominance of state and local control in American education. The general view throughout the executive branch is that the federal government should make no effort to impose a national curriculum in American schools. It is debatable, of course, whether national tests would lead to a national curriculum, but many believe it would, and that charge would surely be made if the federal government were to sponsor a national test. A contrary trend is represented by

Senator Pell's effort to promote a national test. The general feeling, however, is that there is too much opposition in Congress for legislation of this type to pass.

2. The present administration, like others before it, wishes to encourage diversity in the American education system, and the feeling is that this is not compatible with a national test.
3. Since diversity is not only something to be encouraged but, in a major way, is already a fact of American education, the general belief is that no single test would be appropriate for all the curricular variations and emphases that presently exist in American schools.

In sum, we think it is fair to say that both the executive and legislative branches of the federal government want to avoid imposing test standards or curricula from Washington, and we do not believe these strongly-held views are likely to change substantially in the near future. It could be argued that the federal government is already in the business of sponsoring a national test through its support of the National Assessment of Educational Progress (NAEP), but we do not think that the general and non-prescriptive support which the federal government gives to NAEP constitutes an exception to this general policy of avoiding national tests and curricula. Further, it should be noted that the content and operation of NAEP are determined by an inter-state body, the Education Commission of the States, and not by the federal government.

The federal government will become increasingly involved in a variety of ancillary but highly important testing activities. These activities will not only increase the use of tests but also, we hope, improve their quality, thereby helping state and local governments to better carry out whatever testing activities they decide to undertake. More specifically, the federal government will act in the following areas:

1. Research on and Evaluation of Tests. The federal government will carry out directly, and support elsewhere, research on and evaluation of tests and testing. A number of projects are already underway in the National Academy of Sciences and the National Institute of Education (NIE). For example, NIE has set up an Office of Testing and Evaluation, and during 1979 will support a study reviewing and evaluating minimum competency testing programs. This study will not only provide information to help educators make choices about minimum competency testing programs but will also provide a guidebook to assist educators in implementing minimum competency testing programs. The impact of minimum competency testing programs on a variety of student and school outcome measures will also be evaluated.
2. National Assessment of Educational Progress. The federal government will continue to provide support for the National Assessment of Educational Progress, but the content of NAEP tests will continue to be determined by non-federal people and organizations.

3. Program Evaluation. Congress will almost certainly continue to increase its demands for more and better evaluation of the effectiveness of federal education programs, and this in turn will mean an increase in the use of various kinds of tests.

4. Technical Assistance. There will probably be an increase in technical assistance to state and local educational agencies. This assistance will probably be of the type represented by the newly established Title I Technical Assistance Centers in which the Office of Education, following the congressional mandate to upgrade the quality of state and local evaluations of Title I of the Elementary and Secondary Education Act, developed three evaluation models and set up ten technical assistance centers around the country to help states implement these models. Much of this technical assistance goes beyond the problems of evaluation design to the question of how tests should be selected.

5. New Legislation. A bill presently before the Congress authorizes grants to state and local educational agencies for implementing or designing minimum competency testing programs. However, the bill states that "nothing in this section shall authorize the Commissioner (of Education) to impose tests on state and local educational agencies, and no such agency shall be compelled in any way to apply

for funds under this section." This bill also authorizes financial assistance to state and local educational agencies to implement basic skills achievement testing programs for students in grades K through 12.

6. Data Burden Reduction. In noting all these activities, which we believe will have the effect of increasing the use of tests, we should not overlook the fact that the federal government will also be trying to reduce unnecessary testing through its efforts at generally reducing data burden. As we all know, in recent years there has been a widespread reaction to the imposition of new data requirements not only by the federal government in particular, but also by other levels of government and private research. Thus, we believe that while many federal initiatives and activities will lead directly or indirectly to more testing, the federal government will, at the same time, be making efforts to prune and restrict its manifold data requirements in response to the valid public complaint that federally imposed data burdens can get out of hand.

## Summary

As we look to the future, we believe there is bound to be continuing controversy over the use and abuse of tests. Legitimate complaints will continue to be made about the data burden imposed by required testing. Debates will, and should, continue over the questionable validity of many tests. The issue of cultural bias against racial and ethnic minorities will continue to be raised. Teacher unions will continue to oppose the accountability implications of testing programs, and those who are specifically concerned with equity for minorities will continue to point out the larger social effects that selection and admission tests impose upon minorities.

Certainly those who develop and administer tests, and especially those who use test results, must be keenly aware of these valid concerns. But the imperfections of tests, serious as they are, cannot constitute a responsible argument for eliminating or cutting back on the use of tests. Imperfect though they may be, tests are far and away the best means we have for assessing educational needs and performance. As such, tests constitute one of the single most valuable tools we have for improving all aspects of the entire educational process from individual diagnosis to program assessment and resource allocation. It does not, then, seem likely to us that those who would eliminate or curtail tests will prevail. Therefore, it is incumbent on the defenders and users of tests to take seriously and thoughtfully the many valid objections raised against testing abuses if such abuses are to be corrected and the increased use of tests is to be defended.

To sum up, both the federal government and the states will increase their use of tests and their involvement in test-related activities, but the respective roles of these two levels of government will be different. States will expand their use of tests into old and new areas, while the federal government, avoiding any effort to sponsor or impose national tests, will increase its supportive activities to assist the states in their endeavors. If both parties carry out their complementary roles, and if the limitations of tests are understood and their abuses minimized, the increased use of tests that we foresee can make a major contribution to the improvement of American education.

# SUMMATION

## THEMES AND COUNTERPOINT

Michael Scriven
Director
Evaluation Institute
University of San Francisco

The great difficulties in using test results to improve both educational programs and the advising of individuals has been a major theme of this conference. This theme, as particularly expressed by Lee Cronbach and Robert Stake, is certainly relevant to the use of aggregated test results in general. We should note, however, that the caution urged by these testing authorities in the use of test results is in strong tension with another major theme of this conference: that testing, because of its general utility, will continue to be an important element in education. In considering the potential opposition of these themes, we must remember that there have been other occasions on which educators have been excessively cautious about the way in which practical instruments have been used, and that these instruments, despite such fears have been used to the great benefit of society. Are the fears expressed here just one more case of academic pussyfooting and nitpicking--or do they emanate from genuine insights?

Although many of the negative comments made with respect to the interpretation of present tests are indeed valid, we should remember that substantial improvements in testing and in the interpretation of test results are possible, although difficult, and that such improvements can lead to useful instruments. Sometimes the backlash against tests, for example, the IQ tests, exhibits more fallacies than the oversimplified use of tests which produced the backlash. We need to be very cautious about these swings in fashion, just as we need to be cautious about dismissing the advice of these experts who are no strangers to practical interpretation. What changes then are needed in the interpretation and use of tests results? Let me mention one or two points that deserve some attention as we look at the general context of testing.

When the issue of functional literacy began to receive some serious attention, it was not long before we discovered that it was possible to prostitute the concept very quickly, just as in the case of competency based testing, programmed texts, and a dozen other educational fads and fashions in this century. The functional literacy test is a case of criterion referenced testing, and criterion referenced testing can only have merit if it allows us to make straight-forward, valid, and useful interpretations. It was not long, however, before we found that criterion referenced tests of literacy were, in fact, designed to lead to diagnoses whose utility is wholly dependent upon particular theories of reading, etc. Now a good functional literacy test has its items tied to simple real-world literacy needs. It is not very hard to see the significance of the fact that a child cannot recognize the meaning of the word "poison" on a bottle label. But it is not at all clear whether we have a useful criterion referenced test if it tells us something about a special subclass of attack skills, for example. The crucial question is whether diagnoses in terms of attack skills lead us to faster remedies than diagnoses in terms of other taxonomies. If they do not, then we have a criterion referenced test in name only and are in no way nearer to either knowing something useful about the student's literacy level or improving it. Therefore, we are not in possession of anything that can be used to evaluate a reading program. We are not yet very good at looking beyond philosophical presuppositions in our attempts to develop testing categories that are useful and not just formal.

Let me pick up on another point that was missing from most of the conference discussions. Testing continues to be viewed as something apart from teaching. It is an insane view, guaranteed to force a choice that should never be made. If testing is to be useful, the testing community must acknowledge that testing is an integral part of teaching. Teaching without feedback is incompetent teaching. A pre-test/post-test arrangement in a particular class is not just an indispensable component in the evaluation of teaching (both formative and summative), it is the only practical way to tailor the teaching to the needs of the students. And it is one of the few ways in which we can inform students that we care about their curriculum needs and, at the same time, provide ourselves with "instantly interesting topics" for discussion in the early part of the course, namely, topics on which students have already made a commitment and have spent some time. One might even add that pre- and post-testing, together with some explanation of how it is to be used in evaluating the teaching and the course, can make students feel that they have a part in controlling and rewarding good teaching. Much of the objection to testing can be completely eliminated by a sustained, repeated, and adequate explanation of its essential function.

A third general change that we must make concerns what we might call "the education of the populus" about testing. Of course, a great deal of time is spent on this already. Nevertheless, it was clear from some of the discussions with representatives of the press and of government bodies at this meeting that even those in somewhat favored positions could benefit from considerably more in-service education about testing matters. If we view the media as the principal in-service education carrier for the population as a whole, then it becomes terribly important to devote a great deal of time to clarifying what we are doing in testing to representatives of the media. The Association of Education Reporters has made one or two brief efforts in this direction. But more work is needed. It is not enough to provide the media with test results; we have to direct its efforts, and inviting its representatives to this conference is a very good first step.

A simple example of this kind of fallacy that persists but could never survive an hour of intensive instruction concerns the idea that it is a special failing of norm referenced tests that they condemn half those tested to be below the fiftieth percentile. There may be a good reason for viewing norm referenced tests in this way; but every criterion referenced test is going to be normed the minute it becomes widely used (as many of them have already been), and we are going to face tremendous disillusionment when we discover that exactly half the youngsters are once more to be found below the median. To solve some of the reporting difficulties, ETS, ACT, and CTB/McGraw-Hill could get together and issue a quarterly newsletter for media people containing straight-forward and acute pro and con arguments about current issues. How many of the media people, for example, are aware of the studies by Reginald Jones and others which suggest that the whole story about the terrible results of labeling or classifying kids in school is mythological? Our local representatives, judging from their discussions here, are not aware of these studies and continue to perpetuate the myth.

I am now going to descend to a list of comments without any overall structure, because no natural grouping occurs to me, even upon reflection.

Some security problems have emerged and were commented on, especially on minimum competency testing. My grapevine information from Florida suggests that the enormous improvements in success rates when those who failed the first time around retook the tests were due to ingenuity rather than increased competency. There were problems about the size of the pool, and after a while, if you got help with the questions that you did not get right, you would eventually run into them again when you retook the test. There were other less attractive procedures used for getting people through, as one might expect. It is of the utmost importance that we rapidly move to such a large item pool that it can be made public or at least reasonably accessible without any problems of teaching to the test. The right to modify some of the numerical variables or descriptors in the items should be retained, but beyond that one might say that safety lies in numbers.

I was surprised that the Serrano-Priest case was not mentioned at all at the sessions. In a way it represents the most savage comment on all educational research that the courts could make. The judge spent--if I recall the figures correctly--six months in trying to identify some variables other than money to which he could appeal in order to define equality of educational opportunity. He was not able to find any, and I believe that is not because there are none but because we have not worked hard enough to identify them. (They are not by any means easy to identify.) Should the testers spend more time looking at independent variables and providing instruments that will measure those, particularly independent variables that are demonstrably relevant to educational success? In the work of Wiley and others, we have some excellent leads and it may be that testing of input variables should be thought of as a major new thrust of the testing organizations, to parallel their efforts with "dependent" variables.

We heard some suggestions that it was time people looked into "non-traditional" modes of measurement, in particular something other than pencil and paper tests. This strikes me as another piece of mythology, of "wouldn't it be nice if...." thinking. There is a very long tradition of looking at performance testing other than paper and pencil, and to the best of my knowledge it turns out to be a) very expensive, b) usually very time consuming, c) very difficult to validate, d) productive of test anxiety to at least a comparable level, in most cases. Here again is a good topic for better education of our taxpayer and media colleagues. We are just not taking seriously enough the notion of accountability or--to put it bluntly--self preservation. You do not preserve yourself from a taxpayer revolt by publishing validation studies, or invalidation studies for that matter, in scholarly journals or anthologies--or reading a paper at a conference.

"Learner verification" came up--the procedure, now mandated in Florida and elsewhere, that requires data on students' successes in learning from textbooks to be gathered and published and used as a basis for improving the texts. There are difficulties with this, but it obviously has great interest for publishers, either as a threat or as a basis for rewarding a type of sensible R and D approach that normally cannot be cost-justified. Does it involve a dramatic delay in updating books? How costly is it? How well can it be done by a publisher as opposed to an evaluator? Good questions arise about it and we need to be moving pretty hard to settle them, especially for a publishing company that has a testing arm as does McGraw-Hill.

Still hammering away at communications and representations, it seems to me we badly need to simplify and highlight the kind of evaluative data we have on curriculum materials. Test result gains, where the testing is appropriate and independent, can be very useful for this, but where would you find such results? They are not usually to be found in the so-called Educational Materials Review Center of USOE, which relies on book reviews for its evaluations. But if there was a place where such materials as result from serious learner verification activities could be deposited and accessed, they might come to be valued more than book reviews or abstract educational research. Some years ago, I argued that we might be able to replace ERIC by something that I called ELMR, with a notable cost saving and a probable effectiveness gain. ELMR stands for Evaluations of Learning Materials Resource, and it involves quite an old-fashioned hardware component of a four-drawer file cabinet, containing files on the few hundred most commonly asked about new materials and most widely used old materials (methods, etc.). This is a file folder for each of these entries and it contains a description of the program in neutral language (not necessarily the publisher's language), plus the results of any evaluation of the material that are available, graded according to the degree of independence and technical competence of the evaluations. This would normally not take more than both sides of one sheet of paper, so we should have no difficulty in getting a thousand items included in at least a nominal sense, if demand required. The folder on "open classrooms" or "year around schools" or "programmed texts" would, of course, take up a little extra space, though it would mostly involve summaries of several reports plus a cross indexing entry. In any case, ELMR is an important part of a shift in emphasis from mere testing to evaluations preferably based on appropriate testing. Of course, there is a whole other dimension of testing which is related to prescriptions for individual student learning paths, but, as we all know, that could stand some improvement too. The function of ELMR is to stress the way in which other positive results can emerge from testing and to get away from the labeling anxiety that the individual-student testing programs have tended to generate.

One other type of data aggregation that is giving us a good deal of trouble at the moment concerns racial differences. One might argue that one should continue to collect race data in order to be able to identify discrimination in the schools. But one could only make that identification if one could assume an absence of discrimination outside the schools and the complete parity of other entry skills, neither of which seems plausible. Furthermore, one cannot use racial differences in successes as a sign of bias of the tests so it is not clear what good gathering such data does. It certainly causes trouble and perhaps adds unnecessary fuel to racist positions. After all, the mission of the schools is to try to do the best that is possible with each individual youngster, and that should be done independently of race. It is entirely appropriate that studies be done on the input end to make sure that there really is an equal effort being made for everybody with certain needs, but doing tests on the output end is not the way to get answers.

Which brings us to the minimum competency topic, one which received a good deal of comment from various speakers. Clearly, Chris Pipho made a telling point when he noted that public support appears to be for minimum competencies and not necessarily for minimum competency testing. It is a common supposition that the latter either leads to the former or is a necessary part of it, although this might not be the case. Haney made a good point about public schizophrenia when he stated there was considerable hostility and skepticism about testing in general, yet the public is favorable to minimum competency testing. Pipho also stressed the very general structuring of opinion into two groups--the educators being against minimum competency testing and the state legislators (and, as one could see at this conference, state and local agency officials, plus test publishers) being for it.

There are a number of identifiable difficulties with minimum competency testing, as it now stands, but I cannot help feeling that there is an enormous difference between this and competency based teacher education, our most recent fiasco along these lines. There are, after all, no clear research findings about what competencies, other than subject matter ones, are appropriate for teachers in classrooms. But we do not need to rely on research in order to judge what competencies are required for graduates of the schools who wish to go into certain job tracks. It does not take serious research to discover that typing skills below 20 words a minute are not a competency level for people who undertake typing, and that an incapacity to read the language of bank contracts, job ads, and the usual forms that one has to fill out for credit or income tax represents an extremely serious handicap in the world of words. It is perfectly true that no sharp cutting line could be justified, even on pragmatic grounds. Hence, it may be appropriate to move to something like Ellis Page's "conversion tables," published on the back of the graduation certificate, with the competency test scores on the front. The tables would enable the employer to see where on the spectrum of a competency test this individual fell; but we would not withhold graduation since it does represent at least a measure of time served and that may not be entirely irrelevant to working performance. Also, to withhold graduation may be to apply the penalties at the wrong place, and somewhat too late.

Still, we must absolutely resist the absurd suggestion that minimum competency testing is to be judged on whether or not it produces improvement in the educational system. Minimum competency testing is a device to tell people the truth about what is happening. What they do about the truth is up to them; one can hardly complain that it is a bad idea to have the newspapers tell the truth about the present state of the economy because it seems to have no effect except to depress people. The first step towards getting something done is to reveal how much needs to be done. One does not improve the situation by killing the bearer of bad tidings.

It should not be supposed that the results of minimum competency testing will lead us to a clear decision about the fault of the teacher, the school system, the parents, the child, or any other of the myriad factors that go into the result. Although it might be possible in principle to do microevaluation that would lead to such "factor culpability" results, we lack both the means and the financing and any particularly good reason for doing it. We are in the position of someone who has discovered through good diagnostic procedures that their child has a certain disease of a very serious kind; it is not important for the parent to discover the exact historical cause of the disease, the only question of interest is whether the available remedies can produce a significant improvement.

As we get into this matter of who is to do what, it is appropriate to look at the question of who is to do the testing. It is worth remembering that the move to teacher-constructed tests led--expectedly--to poor results. Not only do teachers lack the technical skills to construct serious test batteries, field tests, etc., but they have very mixed motivations, including at least some powerful reasons for wanting to set low standards. Any locally-controlled efforts, given that school boards are typically comprised of parents who would like to see their children score high, are likely to meet with similar ends. There is, thus, a very strong reason for maintaining at least state-level testing programs, and since states have the same kinds of problems as local school boards, there is also good reason for maintaining the National Assessment of Educational Progress, and indeed, continuing to upgrade the quality of its testing. All of this has to be combined with a clear emphasis on local options and the distinction--once again--between testing and particular decisions about curriculum content or teaching methods. Of course, where the national and state assessment programs fail to cover particular features of the local curriculum that are seen as very important, locally-appropriate tests (not locally-constructed tests--in most cases) can be added to supplement the picture.

It seems clear that the time has come for a major effort at curriculum reanalysis, so that one can more carefully identify the rationale for both basic and supplementary components. The notion of "education for survival" needs to be reexamined, not dismissed on the grounds that everybody survives in one sense or another. Some survive considerably better than others, and some part of those survival skills are picked up through education. A major publisher like McGraw-Hill, with the collaboration of its testing arm, is in a unique position to readdress the question of "basic education" and the adequacy of our present materials to cover it. It seems clear enough to me that there are a dozen areas in which virtually nothing is provided in the normal school curriculum to assist students in acquiring skills of the very greatest importance in their future school and work careers. I am thinking of a range of skills from emergency cardiac support, through speed reading to elementary financial management.

Another question we should ask is whether testing should be matched to the curriculum of particular schools. It seems to be assumed, by several speakers, that such a match is ideal and should be sought. On the contrary, absolutistic testing across a wide spectrum of skills and knowledge not based on any study of a particular curriculum is the primary requirement, since this is what keeps a curriculum honest. Once the decision has been made that certain parts of the more obvious basic areas can be dispensed with in favor of some locally preferred curriculum, then one can (at least temporarily) drop out that portion of the basic test and replace it with some locally-oriented testing. But the main service that a state or national testing program or test publisher can provide to the citizens of this country is that primary test, the one that tells us about the cost of the tradeoffs to local interests. How else can one identify the excessive costs of local options, or failures of supposed ingenious ways to achieve the basic skills?

I will conclude with a point directed at the future, and also one that is to some degree optimistic. One of the results of the microprocessor revolution is that simple devices for test administration, test correcting, and the integration of results are now technically feasible at a very low cost. We have not seen the electronic descendant of the teaching machine emerge in substantial quantities yet, although Texas Instruments' new talking computer is a remarkable exception. If we avail ourselves of some of the technological breakthroughs in electronics--the only major area in which the price has come down by a factor of ten and the capabilities/ quality has gone up by something like the same factor in recent years--we can achieve some important gains in testing on local, state, and national levels.

To sum it all up is impossible. But I take away a strong sense that we have the expertise to carry testing through to a new era of greater sophistication and greater social utility without greater costs. That really has to be an inspirational conclusion in these days.